Characterisation of Bipolar Parasitic Transistors for CMOS Process Control.

Thesis submitted by
David Wilson
for the Degree of
Doctor of Philosophy.

Edinburgh Microfabrication Facility,
Department of Electrical Engineering,
University of Edinburgh.

March 1992.

Acknowledgements

Throughout the period in which the work which is described in this thesis was performed, I have benefited greatly from the advice and support of many people. It is my great pleasure to acknowledge them here. I would like to thank my supervisor, Professor John Robertson, for his many helpful discussions, for his guidance in the direction the project should take, for his patience, and for his detailed scrutiny of the work presented in this thesis and relevant papers. I would like to thank Dr Anthony Walton, for his help with published material, for technical and modelling discussions, and for his help in proof reading this thesis. I would like to thank my co-supervisor Dr Bob Holwill, for technical discussions. I would like to thank Mr. Martin Fallon, for many useful discussions, and for software support on the E.M.F. VAX computer. I would like to thank Mr. Alan Gundlach, Mr. John Fraser and all the E.M.F. technicians for helping me operate relevant equipment and make full use of the E.M.F. I would also like to thank Mrs. Liz Patterson for her help in typing many internal reports.

I would like to thank my industrial sponsors Motorola, for financial support. I would also like to thank, Mr. Nigel Davenport, Mr. Paul Tuohy and Mr. Derek Morris for technical support at Motorola. I would like to thank my wife, Mrs Rhona Wilson, for her understanding. Finally, I would like to thank the Science and Engineering Research Council, for their financial support throughout the course of my studies.

Abstract

In integrated circuit manufacture, in particular, quality assurance, QA, is increasing rapidly in importance and in this research methods are developed and assessed which will assist with this.

A review of current IC manufacturing is presented and CMOS technology shown to be dominant with BiCMOS seen to be a growth area. The role of Statistical Process Control, SPC, and the need for QA is also reviewed. This thesis addresses the problem and has defined some new techniques for the process control of a standard CMOS process. The approach is a novel one employing the concept of parasitic bipolar transistor test structures as a process control tool for present day CMOS circuits and, even more importantly, for BiCMOS devices.

Test chip design and manufacture for the project are presented and the techniques proposed include:

a) characterisation of parasitic JFETs to provide well depth information electrically

b) the use of parasitic lateral bipolar transistors to estimate the sideways diffusion component associated with MOS transistors fabricated in a CMOS process

c) the use of parasitic bipolar test structures to evaluate CMOS process uniformity.

The test structures are shown to give a wealth of extra information and provide useful parameters for process control and, in some cases, have even been demonstrated to be more sensitive to CMOS process non-uniformities than those extracted from MOS devices themselves.

The adoption of the concepts presented in this thesis will provide process control information for today's CMOS processes and an insight into the control of future BiCMOS processes.

Contents

1 CMOS - An Introduction **1**
1.1.1 CMOS Technology Past and Present 1
1.1.2 Advantages of CMOS 2
1.1.3 CMOS Fabrication 2
1.2 Future Trends 5
1.3 BiCMOS
The Future Partnership 5
References 7

2 Bipolar Device Theory **9**
2.1 The pn Junction 9
2.1.1 A Thought Experiment 10
2.1.2 The Depletion Approximation 12
2.1.3 Linearly Graded Junction 17
2.1.4 The pn Junction Under Forward and Reverse Bias 18
2.1.5 Current Voltage characteristics of the Ideal Junction 22
2.1.6 Generation and Recombination 25
2.2 Principles of Transistor Operation 26
2.2.1 1-D Analysis of npn Device 26
2.2.2 Non Ideal Factors 33
2.3 MOS Device Operation 33
References 43

3 Bipolar and MOS Modelling **45**
3.1 Parametric Test Systems 46
3.1.1 Tester uses 46

3.2 Parametric testing for Statistical Process Control (SPC) 47
3.2.1 Six Sigma Capability 48
3.2.2 Statistical Process Control 48
3.3 Bipolar modelling 50
3.3.1 The Ebers and Moll equations 50
3.3.2 The EM2 model 57
3.3.3 The Gummel Poon Model 61
3.4 The MOSFET model 66
3.4.1 The Level 1 Model 67
3.4.2 The Level 2 Model 67
3.4.2.1 Effective Channel Length 68
3.4.2.2 Threshold Voltage 68
3.4.2.3 Effective mobility 69
3.4.2.4 Saturation Voltage 70
3.4.2.5 Channel Length Modulation 70
3.4.2.6 Channel Shortening at Punch-Through 71
3.4.2.7 Mobility and Channel Modulation 71
3.4.2.8 Drain Current Equations 72
3.5 TECAP Operation and Hardware Requirements 72
3.5.1 Hardware Requirements 73
3.5.2 TECAP Models 77
References 78

4 Test Chip Design, Fabrication And Characterisation 80
4.1 Aims of The Test Chip 80
4.2 The EMF 5μm Process 81
4.2.1 Design Rules 82
4.3 Initial Design Thoughts 87
4.3.1 Traditional bipolar and JFET designs 87
4.3.2 Feasibility with EMF 5μm process 92
4.3.3 Initial device simulation 93
4.4 Test Chip Design 94
4.4.1 The Lateral Bipolars 99
4.4.2 The Verticals 104
4.4.3 The JFETs 104
4.4.5 The MOSFETs 110
4.4.6 The Layout 110

4.5 Well Depth Process Split 110
4.6 Other Applications 111
References 114

5 The use of parasitic JFETs to monitor Well depth 115
5.1 Introduction 115
5.2 JFET electrical theory 115
5.3 Device Analysis 117
5.4 JFET sensitivity to n-well depth 126
5.4.1 Characterisation model 126
5.4.2 Initial characterisation 131
5.4.3 JFET Threshold Measurement 131
5.4.4 Junction Leakage currents 131
5.5 Relationship of n-well depth to JFET threshold 139
5.6 Cross wafer variation 142
5.7 Conclusions 143
References 147

6 ΔL Extraction Using Parasitic Bipolar Transistors 148
6.1 Introduction 148
6.2 Theory of Lateral Transistor Operation 149
6.3 ΔL Extraction 154
6.4 Simulation 154
6.5 Experimental Method 157
6.5.1 Device Processing 157
6.6 Electrical Characterisation 157
6.6.1 R_E and R_C Extraction 158
6.6.2 V_{AF} and V_{AR} Extraction 158
6.6.3 Forward Gummel Characterisation 169
6.6.4 Reverse Gummel Extraction 169
6.6.5 Parameter Optimisation 173
6.6.6 Optimisation and Process Control 175
6.6.7 Structural Considerations 175
6.6.8 ΔL Extraction 179
6.6.9 Simulated vs Measured Results 179
6.6.10 Sensitivity to V_{EB} 180
6.6.11 Significance of K 189

6.6.12 Comparison With Conventional Methods 189
6.7 Conclusions 189
References 193

7 CMOS Process Uniformity Evaluation Through the Characterisation of Parasitic Transistors 194
7.1 The Vertical Parasitic Bipolar Transistor 195
7.1.1 Device Structure 195
7.1.2 Characterisation Technique 196
7.1.3 Parametric Results 196
7.1.4 R_E and R_C 196
7.1.5 Early Voltage 204
7.1.6 Forward Gummel Parameters 204
7.1.7 Reverse Gummel Parameters 217
7.2 The MOS Transistor 218
7.2.1 Device Structure 218
7.2.2 Characterisation Technique 219
7.2.3 MOS Parameter Variation with Process Split 225
7.5 A Process Uniformity Evaluation 226
7.6 Conclusion 239
References 241

8 Conclusions and Further Work 243
8.1 Conclusions 243
8.2 Further Work 245

List of Symbols

E	electric Field
C_{ox}	oxide capacitance
C_{OX}	oxide capacitance scaled by mosfet dimensions
$D_n,\ D_p$	electron, hole diffusion coefficient
G_o	conductance
I	current
I_0, I_S, I_i	saturation current
$I_{0,gr}$	recombination current
I_D	emitter defect current / drain current
I_E, I_B, I_C	emitter, base, collector current
$I_{KF},\ I_{KR}$	knee currents associated with gummel-poon model
J	current density
K_P	intrinsic transconductance parameter
L_{eff}	effective channel length
$L_n,\ L_p$	electron, hole diffusion length
n	electron density
n_i	intrinsic carrier concentration
n_0	equilibrium electron density
$n_0(x),\ n(x)$	electron density at point x
n'	difference between equilibrium and actual electron density
n_{oN}	n type equilibrium electron density
n_{oP}	p type equilibrium electron density
N_D	donor Concentration
N_{sub}	substrate doping concentration
p	hole density
p_0	equilibrium hole density
$p_0(x),\ p(x)$	hole density at point x
p'	difference between equilibrium and actual hole density
p'_{ne}	n side excess hole concentration (emitter)
p_{oN}	n type equilibrium hole density
p_{oP}	p type equilibrium hole density
Q_n	total free charge in channel
$Q_{BT},\ Q_{BO}$	total base charge, normalised total base charge
R_S	sheet resistance
U_0	normalised potential
$U_{crit},\ U_{exp}$	critical field for mobility degradation/ critical field exponent
U_{tra}	transverse field coefficient
$V_A,\ V_B$	forward and reverse early voltages
V_G, V_D, V_S	gate, drain, source voltage
V_T, V_{TO}	threshold voltage, zero bias threshold voltage
W_{bo}	drawn base width
$x_{\mu N}$	n side boundary position of a junctions space charge region
$x_{\mu P}$	p side boundary position of a junctions space charge region
X_0	space Charge Layer thickness
ϕ_i	built in potential
ϕ	surface potential
ψ_N	n-side Potential
$\psi(x)$	potential at point x
$\Delta\psi_0,\ \psi_{BI}$	contact potential
λ	channel length modulation parameter
ρ	resistivity
μ_{eff}	effective mobility
μ_0	surface mobility at low gate voltages
$\mu_n,\ \mu_p$	mobility of electrons / holes

Chapter 1

CMOS - An Introduction

1.1.1 CMOS Technology Past and Present

After its inception in the early 60's [1-2] CMOS logic was restricted in its applications to very specialised circuitry. These circuits were incorporated into watches and calculators and other low power or radiation-sensitive devices. In the years after its debut CMOS was compared unfavorably with NMOS due to its lower speed and poorer packing density. It was also more complex in it's layout and special attention had to be given to latch-up prevention.

In the 1980s transistor count passed the 4 million mark on the most complex designs. With this increase in transistor density, power dissipation has become a fundamental limit to circuit design. With the push for smaller geometries in the mid 1980s, CMOS performance began to align itself with that of NMOS technology. Small p-channel devices now deliver current approaching that of their n-channel counterparts. As devices are scaled below 1μm gate length, the difference between hole and electron saturation velocity decreases [3-4]. CMOS DRAMs with access times under 70 ns are commonplace.

As the MOS technology was scaled down, NMOS processing has evolved to more complexity to provide punch-through protection, multiple threshold voltages, buried contacts and polysilicon interconnect and or resistors.

Conversely, the complexity of CMOS processes has remained relatively unchanged and perhaps even improved with various process innovations. CMOS now has the largest share of the semiconductor device

market (64%) and the growth in the CMOS market throughout the 80s has been almost entirely at the expense of NMOS; figure 1.1. When compared with NMOS, CMOS now has much lower power dissipation, comparable speed, and slightly poorer packing density, but better reliability. Packaging constraints will continue to limit chip power dissipation to a few watts per chip in order to avoid reliability issues [5].

1.1.2 Advantages of CMOS

Several of the dominant features of CMOS technology described in ref [6] are summarised below

1. Low power. DC conducting paths to ground do not arise: current is only drawn during switching. Reduced power means less strain on wire bonds and metal lines.

2. Higher noise margins. The maximum low output voltage is about 2 volts, compared with 0.5 volt for both TTL and NMOS.

3. Ability to withstand extreme temperatures. This has led to CMOS being utilized for engine management chips. Thermal runaway is not a problem. Like any MOS IC, CMOS is self limiting: circuit currents drop as temperature increases.

4. Sharp transfer characteristics. Output voltages switch quickly in response to input voltages.

5 Wide supply tolerance.

6. Bipolar transistors and analogue circuits can be constructed on the same chip as CMOS logic. This is discussed later.

1.1.3 CMOS Fabrication

Conventional CMOS was formerly realized by putting n-channel transistors in a p-well formed by diffusing boron atoms into an n-type substrate. p-channel devices were made outside the well in the n-substrate. Figure 1.2 shows the cross-sectional view of a CMOS

structure made by the p-well technology. The start of the 80s saw this approach reversed with p-channel devices formed in an n-type tub [7][8]. This structure is identical to that in figure 1.2 with the n and p labels interchanged. Other working alternatives are twin-well [9], retrograde well [10, 11] and quadruple well technology [12]. The choice of technology chosen depends on the specific applications. Some of the pros and cons of the various well configurations are given below.

p-well vs n-well

Present CMOS technology offers sub-micron design rules. At these dimensions, n-channel devices in a p-well, when compared with p-channel devices in a n-well, provide about twice the driving current [13] but almost four orders of magnitude higher substrate current [14]. This is a fundamental difference between the two device types and it affects technology selection in many areas. Another fundamental limit is that the doping concentration in the well has to be higher than that in the starting substrate, thus resulting in higher junction capacitance and more body effect for devices made in the well. It has been shown [15] that, due to better intrinsic gettering, a p-type epitaxial layer grown on a p+ substrate provides longer (millisecond) minority lifetime than does a n-type epitaxy on n+ substrate.

Twin well

The twin well approach forms two separate wells for n- and p-channel transistors in a lightly doped substrate. The major advantage of the twin well approach is the flexibility of selecting a substrate type (n or p) with no effects on transistor performance. The latch-up behaviour, however, will not be identical. A description of the latch-up problem and the influence of substrate choice is given in [16]. This flexibility may be important in implementing designs with different applications and in addition, self-aligned channel stops can be easily implemented with this approach. Consequently, spacing between n- and p- channel devices can be reduced for high density circuits. As CMOS technology advances to half micron dimensions, the twin-tub approach has become more attractive for the following reasons. Because the two device types

perform similarly in the half micron regime, it makes sense to provide symmetric n- and p-channel devices [17]. Since the doping concentration has to be scaled up at these dimensions, whether the devices are made in the well or substrate makes only a marginal difference. Submicron technologies such as trench isolation and epitaxial substrate, work well with the twin-well approach. For example, trench sidewalls are butted against highly doped wells. Moreover, when epitaxy is used, this approach offers great flexibility in choosing n-on n+ or p-on p+, and even n- on p+ or p-on n+ if BiCMOS (bipolar/CMOS) chips are implemented.

Retrograde Wells

Conventional wells are formed by diffusion. This is an isotropic process where impurity atoms diffuse laterally as well as vertically. Lateral diffusion takes up silicon area resulting in poorer packing density. High energy ion implantation, when used for well formation, provides minimum lateral spread because of the anisotropic behaviour of the implantation process. Unlike a diffused profile in which peak concentration is always at the silicon surface, the peak of the profile is buried at a certain depth (depending on the implant energy) inside the silicon surface and the impurity concentration decreases as it approaches the silicon surface. This type of profile is called a retrograde well as it has the advantages of

(i) providing a retarded electrical field, which reduces the vertical bipolar gain (the properties of vertical bipolar gain with respect to well profile is discussed in chapter 7);

(ii) high conductivity at the bottom of the well which decreases voltage drop in the well and increases vertical punch-through when the well is made shallow;

(iii) reduced junction capacitance and body effect if the implant energy is sufficiently high to move the highly doped region away from the channel.

For BiCMOS applications, the highly conductive layer near the bottom of the retrograde well can also be used as a buried layer if bipolar

devices are made in the well.

1.2 Future Trends

With its inherent low power characteristics and advantages in circuit design, CMOS is undoubtedly the dominant VLSI technology. In the next decade most digital designs, including microprocessors and memories will be made using CMOS. NMOS technology will disappear almost entirely (as did PMOS). More bipolar applications will be switched to CMOS as the technology becomes the dominant analogue technology. In ULSI, only CMOS circuits can run very fast without getting very hot. BiCMOS cells will be present in an increasing number of CMOS designs, to provide very high speeds and large drive currents. 16Mb DRAMs will soon be widely available and working 64Mb DRAMs have already been demonstrated [18]. Two level metal interconnect is already common with three and four level interconnect a necessity [19]. New technologies such as three-dimensional IC's will be more effective in packing more transistors on a chip or achieving a higher level of integration. CMOS due to its low power advantages will be the dominant technology in three-dimensional ULSI circuits.

1.3 BiCMOS: The Future Partnership

Figure 1.1 predicts a 5% market share for BiCMOS in 1994. This will probably be an underestimate as bipolar cells will be used in CMOS circuits without being regarded as a fully integrated BiCMOS process. The work presented in this thesis explores the concept of parasitic bipolar transistors as a process control tool for present day CMOS circuits. The advantages of this technique would be twofold.

(i) Bipolar test structures fabricated in a CMOS process can provide a wealth of extra process information which has so far been untapped.

(ii) It will provide the MOS process engineer with an opportunity to become familiar with the concepts of bipolar transistor operation before it becomes an absolute necessity in

the everyday fabrication of BiCMOS devices.

This thesis will demonstrate simple bipolar device designs that can be incorporated into any CMOS process. The theory of operation of bipolar transistors is developed in chapter 2. MOS device theory is discussed briefly to give a background to the device modeling aspects which are covered in detail for both the bipolar and MOS device in chapter 3.

The results presented in the thesis offer a comparison of conventional CMOS parameters vs those extracted from bipolar devices when used for CMOS process control. This illustrates the advantages of the parasitic device when applied to particular areas of CMOS process control.

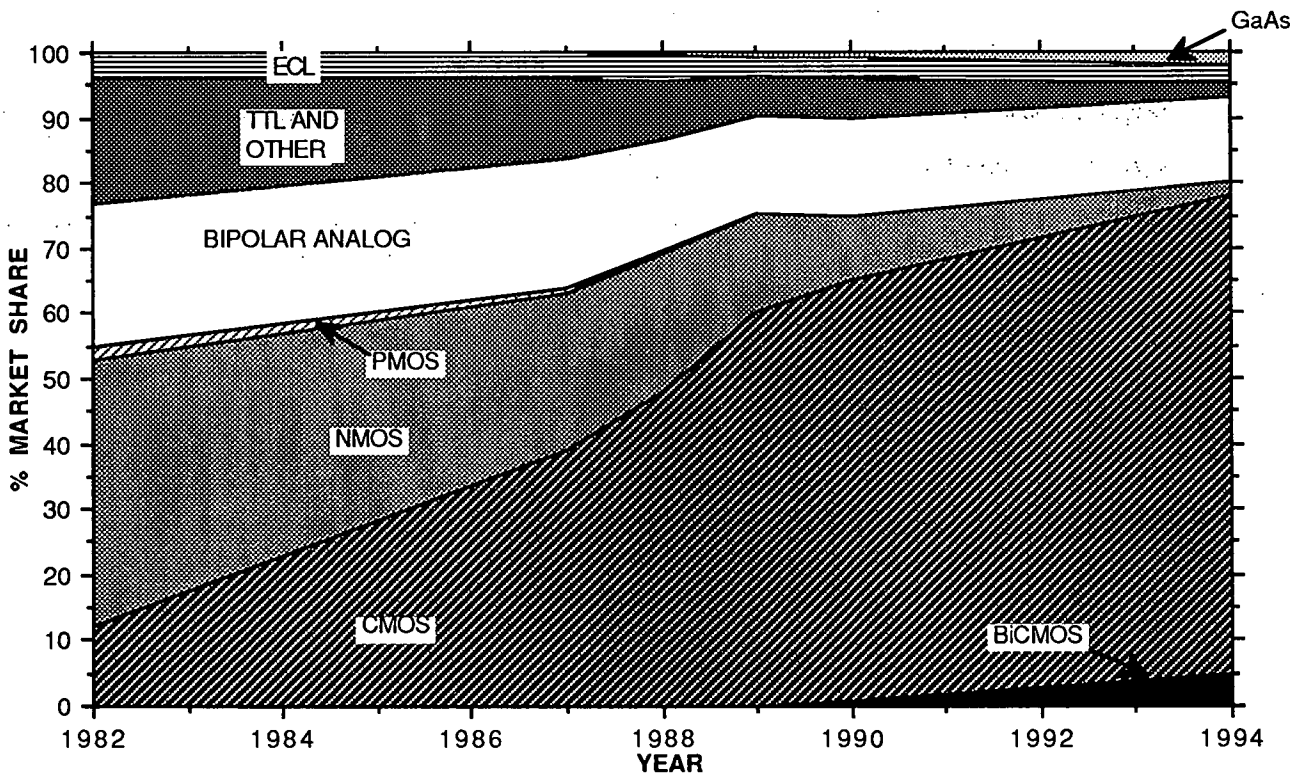


Figure 1.1 Technology market share, including predicted shares.

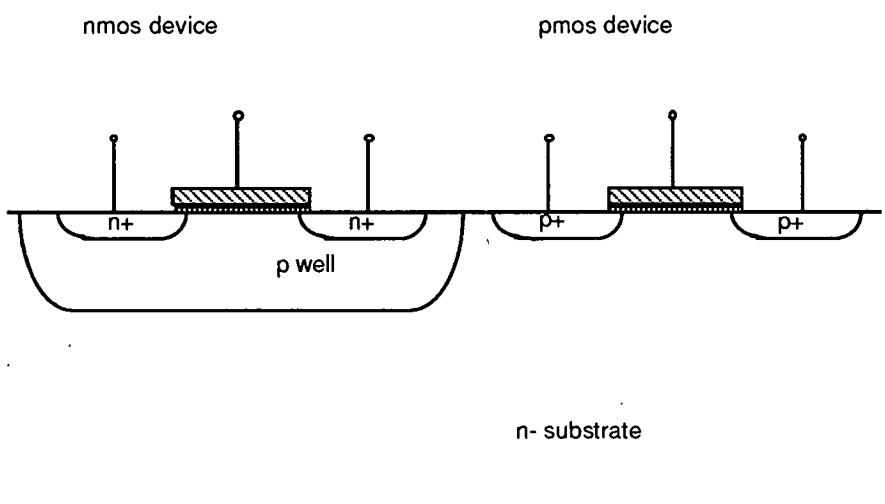


Figure 1.2 Schematic showing main features of a CMOS cell fabricated using p-well technology.

References

[1] F.M. Wanlass and C.T. Sah,"Nanowatt Logic Using Field-Effect Metal-Oxide Semiconductor Triodes,"*ISSCC Digest,32*, February 1963.

[2] G.E Moore, C.T. Sah and F.M Wanlass,"Metal-Oxide Semiconductor Field-Effect Devices for Micropower Logic Circuitry," *Micropower Electronics*, edited by E. Keonjian, Pergamon Press, p.41, 1964.

[3] H. Shichjio,"A Re-examination of Practical Performance Limits of Scaled n-Channel and p-Channel MOS Devices for VLSI,"*Solid-State Electronics*, Vol 26, p. 969, 1983.

[4] A. Schmitz and J.Y. Chen,"Design, Modelling and Fabrication of Subhalf-Micrometer CMOS Transistors,"*IEEE Trans. Electron Devices*, Vol ED-33, p.148, 1986.

[5] R. Davis, "The Case for CMOS,"*IEEE Spectrum*, p. 26, 1983.

[6] A. G. Buttar,."CMOS - the options,"*CASE project report*, University of Edinburgh, 1982.

[7] K. Yu et al,"HMOS-CMOS: A Low Power High Performance Technology,"*IEEE J. Solid-State Circuits*, vol. SC-16, p. 454, 1981.

[8] R. Chwang and K. Yu,"CHMOS:An n-well Bulk CMOS Technology for VLSI,"*VLSI Design*, 4th quarter, p. 42, 1981.

[9] L.C. Parrillo, R.S. Payne, R.E. Davis, G.W. Reutlinger, and R.L. Field,"Twin-Tub CMOS:A Technology for VLSI Circuits,"*IEDM Tech. Dig.*, p. 752, 1980; also, L.C. Parillo, L.K. Wang, R.D Swenumson, R.L. Field, R.C. Melin, and R.A. Levy,"Twin-Tub CMOS II: An Advanced VLSI Technology,"*IEDM Tech. Dig.*, p. 706, 1982; also, J. Agraz-Guerena, R. Ashton, W. Bertram, R. Melin, R. Sun, and J.T. Clemens,"Twin-Tub III-A Third Generation CMOS Technology,"*IEDM*

Tech. Dig., p. 63, 1984.

[10] R.D. Rung, C.J. Dell'Oca and L.G. Walker,"A Retrograde p-well for High-Density CMOS,"*IEEE Trans. Electron Devices*, vol. ED-28, p. 1115, 1981.

[11] S.R. Combs,"Scalable Retrograde p-well CMOS Technology,"*IEDM Tech. Dig.*, p. 346, 1981.

[12] J.Y. Chen,"Quadruple-Well CMOS for VLSI Technology,"*IEEE Trans. Electron Devices*, vol. ED-31, p. 910, 1984.

[13] R.A. Martin and J.Y. Chen,"Optimized Retrograde n-Well for One Micron CMOS Technology,"*Proc., Custom Integrated Circuits Conf.*, p. 199, 1985.

[14] R. Chwang and K. Yu,"CHMOS:An n-well Bulk CMOS Technology for VLSI,"*VLSI Design*, 4th quarter, p. 42, 1981.

[15] J.O. Boreland et al,"An Intrinsic Gettering Process to Improve Minority Carrier Lifetimes in Mos and Bipolar Silicon Epitaxial Technology,"*Semiconductor Processing, ASTM STP 850*, edited by Dinesh C. Gupta, American Society for Testing and Materials, 1984.

[16] R. S. Muller and T. I. Kamins,"Device Electronics for Integrated Circuits,"*Wiley International 2nd ed*, pp458-462, 1986.

[17] S. Kohyama, J. Matsunaga, and K. Hashimoto,"Directions in CMOS Technology,"*IEDM Tech. Dig.*, p. 151, 1983.

[18] Y. Nakagome et al,"An Experimental 1.5V 64-Mb DRAM", *IEEE J. Solid-State Circuits*, vol. SC-26, p. 465, 1991

[19] S. R. Wilson, J. L. Freeman, C. J. Tracey," A Four-Metal Layer, High Performance Interconnect for Bipolar and BiCMOS Circuits",*Solid State Technology*, Vol. 34 No. 11,p 67, 1991.

Chapter 2

Bipolar Device Theory

Introduction

This chapter presents basic device theory for the bipolar junction transistor and the MOS transistor. The theory presented here will be used as a starting point to develop the concept of device modelling in chapter 3. This thesis concentrates on the characterisation of the BJT for CMOS process control. Other devices such as JFETs and lateral BJTs are examined later in the text for distinct purposes. It was felt that as these devices are somewhat specialised, the treatment of their device theory and characterisation should be included in the relevant chapters.

2.1 The pn Junction

A p-n junction is the boundary between a p-doped region and a n-doped region in a semiconductor single crystal. All integrated circuits and almost all silicon devices depend upon the characteristics of pn junctions for their operation [1]. The basic electrical characteristics of the p-n junction were first proposed by Shockley [2]. The initial theory was then expanded by Sah, Noyce and Shockley [3], and by Moll [4]. Two comprehensive but differing views of the development of the most important semiconductor theory are given in reviews by Shockley [5] and Moll [4].

This section will deal with the basic junction theory for the linearly graded junction and the step junction under equilibrium and reverse bias conditions. The depletion approximation will be introduced as well as topics such as recombination. Although the step and linearly graded junctions are only two of many possible theoretical junction types they provide a good illustration of typical junction behaviour.

2.1.1 A Thought Experiment

The basic concept of a pn junction can be introduced by using a thought experiment [6, 1]. Imagine two semiconductor single crystals both uniformly doped, one p-type the other n-type. In these crystals there exists a condition of space charge neutrality. We take these two crystals at room temperature and join them together perfectly, even at an atomic level, to make one crystal. This process is physically unrealistic but can be imagined easily as demonstrated in figure 2.1. Let us consider what happens to the carriers in each crystal. The instant they are joined there is a sharp increase in holes and decrease of electrons going from the n-type region to the p-type region. Figure 2.1 shows this schematically. From Fick's first law we know that such a large gradient (many orders of magnitude) will cause diffusion. Hence, there is a flux of electrons from the n-type region to the p-type and conversely a flux of holes from the p region to the n. This flux of holes and electrons spreads out the abrupt transition in carrier concentration.

However, we know that for an electron to leave the n-side of the junction and enter the p-side it must leave behind a stripped donor atom. These donor atoms are fixed in the silicon lattice and cannot move. Thus space charge neutrality close to the junction disappears as electrons diffuse leaving stripped donors. Similarly holes diffuse from the p-side to the n-side leaving stripped acceptors, (figure 2.1). The removal of space charge neutrality in the junction area causes an electric field to be formed. Conventionally, lines of force originate on a positive charge and end on a negative charge, thus the field lines go from the n-side to the p-side. The presence of such a field causes a drift current of electrons to flow from the p-side to the n-side and holes from n to p.

The thought experiment has now developed four mathematical current components, the electron and hole diffusion and drift currents. If these currents had physical as well as mathematical existence there would be power dissipation as each current component moved through the silicon. However, since energy dissipation is dependent on net current, we resolve this dilemma by having zero net current at equilibrium. This means the hole drift current must match the hole diffusion current with no net hole transport in either direction. This must also be true for electrons.

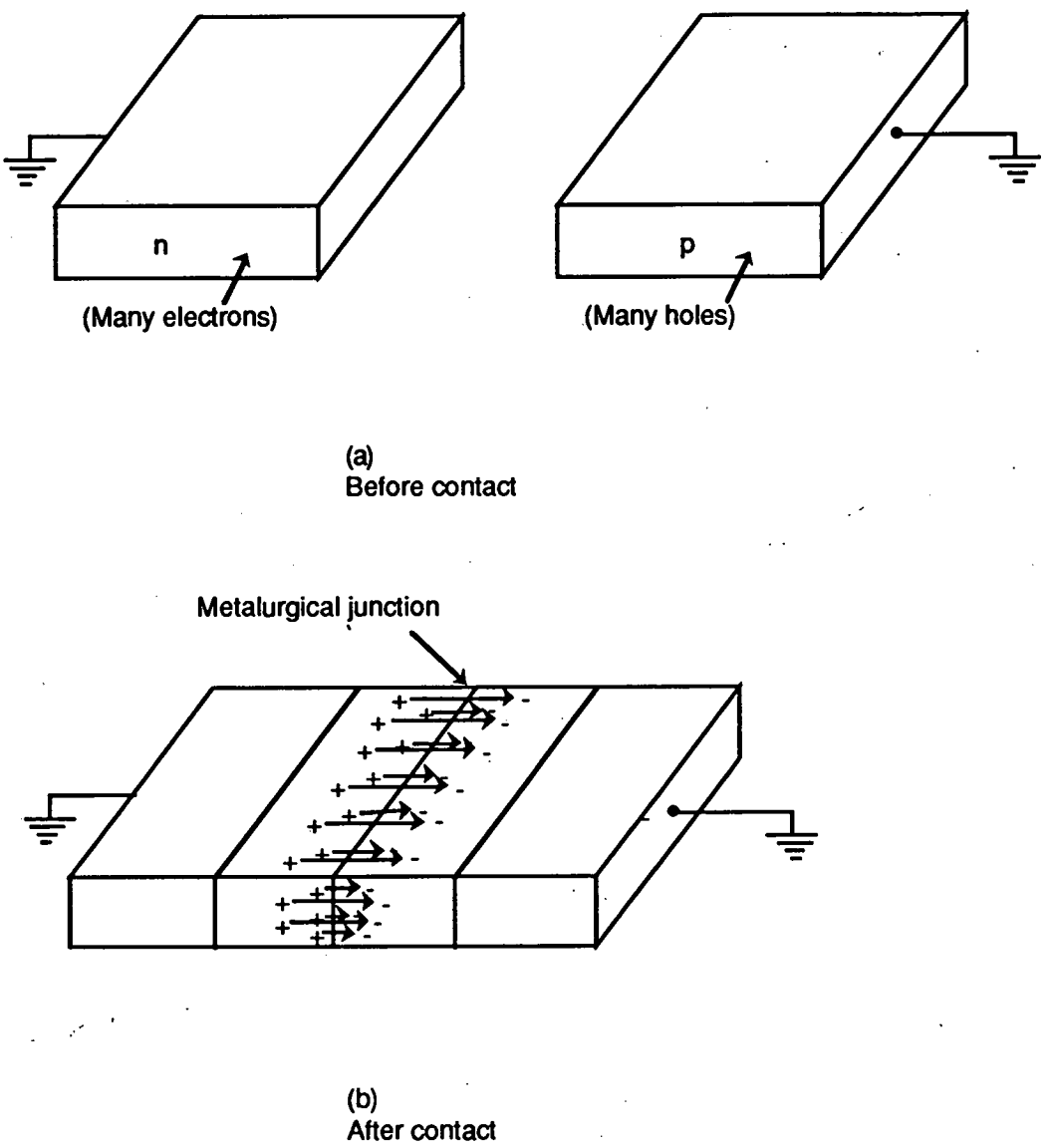


Figure 2.1 Schematic illustration of "thought experiment" showing the "perfect" joining of two crystals at atomic level.

Thus, what emerges from this thought experiment is a perfect crystal which changes type at a planar pn junction. It has a positive charge associated with the n-side and an equal negative charge associated with the p-side. The resultant electric field reaches a maximum at the junction. If we look at hole carrier concentration it is at its maximum, the equilibrium density, far from the junction on the p side. As we approach the junction it begins to fall and well into the n-type material, it has fallen to its equilibrium minority value. There is a symmetrical distribution of electron concentration reflected about the planar junction.

2.1.2 The Depletion Approximation.

In this section we consider the electric field created by the fixed charge associated with the stripped donor atoms in the n-type side of the crystal developed in the thought experiment above. We can examine the field intuitively with the aid of figure 2.2. The closer donor atoms are to the junction the more likely they are to have lost their charge to the diffusion current. Using the concept of field lines we can create a 1-D spatial field distribution shown in figure 2.3. The electric field at any point is proportional to the density of the field lines. The maximum field must be at $x=0$, the junction, as all lines must pass through the junction according to the definition above.

The depletion approximation idealizes the charge density profile shown in figure 2.4b to give that shown in figure 2.4c. The result is a region from $X_o/2$ to $-X_o/2$ where no carriers exist or at least their density is negligible. This depletion approximation is sometimes referred to as the abrupt approximation.

The approximation makes the mathematical manipulation of the junction equations simpler. Using Poisson's equation to deal with the space charge region on the n-side:

$$\frac{dE}{dx} = \frac{qN_D}{\varepsilon} \tag{2.1}$$

Then separate the variables and integrate with the space charge limits on the n-side $-X_o/2 < x < 0$

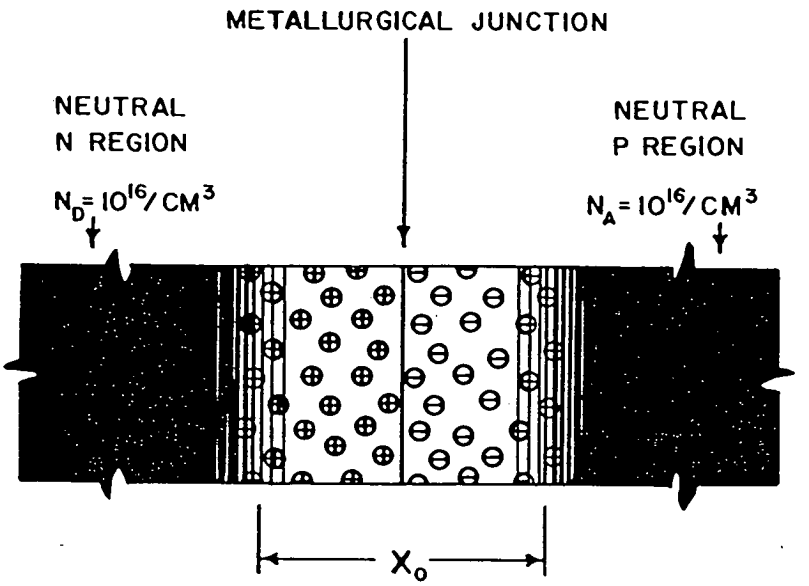


Figure 2.2. Pictorial representation of junction at equilibrium having both sides doped to $10^{16}/cm^3$ with circled symbols depicting ionized impurity atoms. Taken from reference [19].

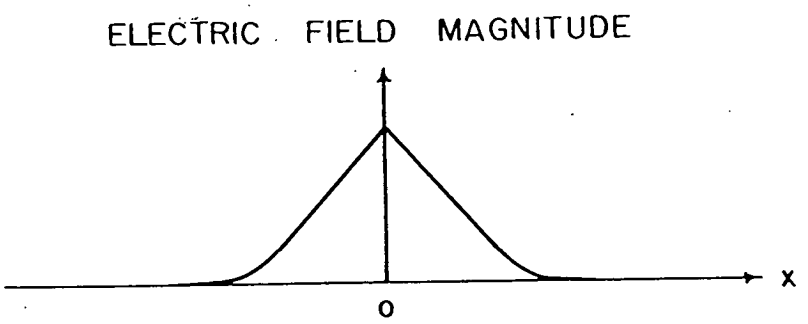


Figure 2.3. Resulting electric field profile from figure 2.2.

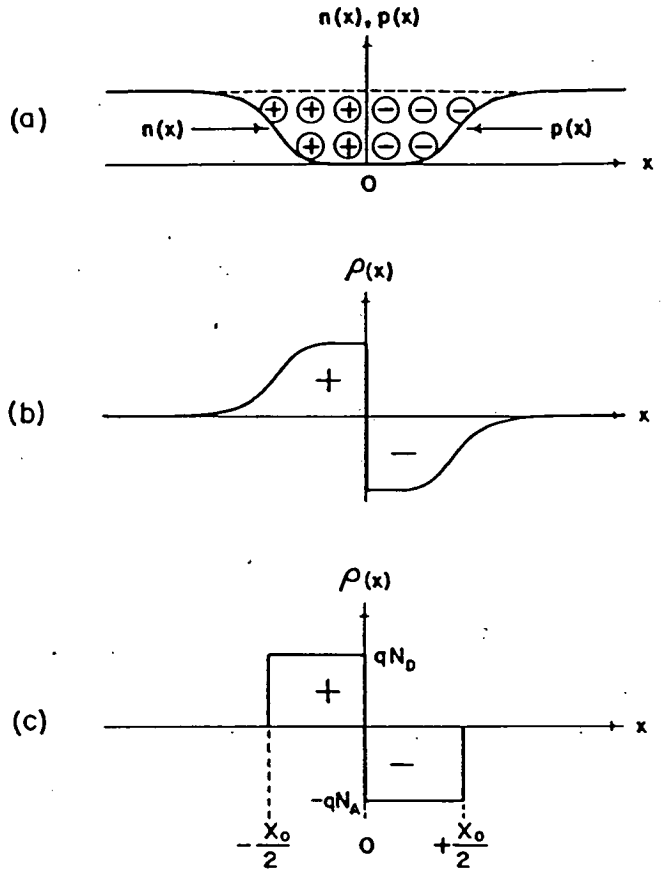


Figure 2.4 Applying the depletion approximation to the symmetric junction of figure 2.2. (a) Carrier profiles with heuristic representation of ionic charges "uncovered" by depleted carrier populations. (b) Charge-density profile consistent with (a). (c) Idealized "depletion approximation" charge density profile corresponding to actual profile in (b). Taken from reference [19].

$$\int_{0}^{E(x)} dE_1 = \frac{qN_D}{\varepsilon} \int_{-X_o/2}^{x} dx_1 \tag{2.2}$$

Using a dummy variable for integration yields,

$$E(x) = \frac{qN_D}{\varepsilon}x + E_m \tag{2.3}$$

where

$$E_m = \frac{qN_D X_o}{2e}. \tag{2.4}$$

To obtain an analytic expression for the electrostatic potential developed on the n-side of the junction, substitute $\frac{d\psi}{dx}$ for $E(x)$ and integrate over the same boundaries as in equation 2.2. Let the electrostatic potential on the n-side be ψ_N, such that

$$\int_{\psi_N}^{\psi(x)} d\psi_1 = \frac{-qN_D}{\varepsilon} \int_{-X_o/2}^{x} \left(x_1 + \frac{X_o}{2}\right) dx_1 \,. \tag{2.5}$$

This gives

$$\psi(x) - \psi_N = \frac{-qN_D}{\varepsilon}\left(\frac{x^2}{2} + \frac{X_o x}{2} + \frac{X_o^2}{8}\right) \tag{2.6}$$

set $\psi=0$ at $x=0$ so that

$$\psi_N = \frac{qN_D X_o^2}{8\varepsilon}. \tag{2.7}$$

Thus

$$\psi(x) = \frac{-qN_D}{2\varepsilon}\left(x^2 - X_o x\right). \tag{2.8}$$

However, in the example chosen for the thought experiment $N_A = N_D = N$, thus the combined electrostatic potential of the p-side plus the n-side is found from equation 2.7 and is twice $|\psi_N|$ or

$$|\Delta\psi_o| = \frac{qNX_o^2}{4\varepsilon}. \tag{2.9}$$

This equilibrium potential is sometimes referred to as the built-in potential ψ_{BI}. Rearranging equation 2.9 results in an expression for the depletion width of the junction at equilibrium.

$$X_o = \left(\frac{4\varepsilon|\Delta\psi_o|}{qN}\right)^{1/2} \tag{2.10}$$

The Boltzman relation can be introduced,

$$\frac{n_{oN}}{n_{oP}} = exp\frac{q|\Delta\psi_o|}{kT} \tag{2.11}$$

substituting $n_{oP} = n_i^2/p_{oP}$ and inverting

$$|\Delta\psi_o| = \frac{kT}{q} ln\frac{n_{oN}p_{oP}}{n_i^2} \tag{2.12}$$

but for our example

$$N \approx n_{oN} = p_{oP} \tag{2.13}$$

so

$$|\Delta\psi_o| = \frac{2kT}{q} ln\frac{N}{n_i}. \tag{2.14}$$

From 2.10 and 2.14

$$X_o = \left[\frac{8\varepsilon kT\, ln\left(\frac{N}{n_i}\right)}{q^2 N}\right]^{1/2} \tag{2.15}$$

Using a "typical" value of N = 10^{16}/cm^3 results in $\Delta\psi_o$ = 0.69V and X_o = 0.43μm.

To summarise this section one can say that the contact potential $\Delta\psi_o$ may be regarded as the potential hill that keeps majority electrons on the N side. That is a steady current of electrons diffuses "over" the hill because of the concentration gradient existing there, but an equal current of electrons in the other direction is maintained at equilibrium through drift caused by the built-in field. The same potential hill $\Delta\psi_o$ will keep holes on the P side. $\Delta\psi_o$ will adjust itself so that with it's accompanying field it is just able to counter hole and electron diffusion at equilibrium.

2.1.3 Linearly Graded Junction

The linearly graded junction is another doping profile that, like the step junction, can be treated exactly. This treatment also gives useful results for approximating real *pn* junctions. In a linearly graded junction the net dopant concentration varies linearly from the p-type material to the n-type material. This type of junction is characterised by a constant a, which is the gradient of the net dopant concentration and thus has units of cm^{-4}. The net dopant concentration can be written as $N_D - N_A = -ax$ throughout the space charge region (figure 2.5a). The field and potential are readily found from Poisson's equation by using the depletion approximation (section 2.1.2). Since the space charge varies linearly with position in the depletion layer, the field varies quadratically and the potential varies as the third power of position in the space charge region (figure 2.5).

Although linearly graded junctions are not realised physically, many practical cases can be approximated by a linearly graded junction over at least a limited voltage range. If a step junction is heated so that the dopant atoms diffuse across the junction, the junction becomes less abrupt. This may be approximated by a linearly graded junction

provided the space-charge region is narrow compared to the diffusion length of the impurity atoms. Even diffused junctions are sometimes approximated by linearly graded junctions over a limited distance.

2.1.4 The pn Junction Under Forward and Reverse Bias

This section will discuss the low level behaviour of the pn junction under forward and reverse bias conditions. Although pn junctions are quite often operated under high level conditions, the low level case allows the concepts involved to be grasped more easily. The term low level here means that the minority carrier densities remain much smaller than the equilibrium majority carrier densities. However, the minority carrier densities can exceed their own equilibrium density by more than several orders of magnitude and still fulfill this condition.

For simplicity we shall continue to use the 1-dimensional sample developed from the thought experiment, i.e., a uniformly doped single crystal with a symmetric step junction. We shall consider only one carrier, electrons. In section 2.1.1 it was established that in equilibrium, at any point in any plane of the model crystal, the drift current and the diffusion current balance perfectly with no net current flow. From 2.1.2 the junction possesses a built-in potential $\Delta\psi_o$. A forward potential will reduce this potential step, and figure 2.6 shows this effect schematically, highlighting the effect on the width of the transition region and the resulting reduced electric field.

This change favours the diffusion component of the junction current while inhibiting the drift component. Consequently, electrons flood from the n-side of the crystal to the p-side (also holes flood from the p-side to the n-side). This injection of electrons forms the basis of transistor action as will be discussed later. It is the quantities of these injected carriers which are treated in this section and lead to the derivation of the low level current equations for pn junctions under bias.

Figure 2.6 shows the sample schematically with a forward bias V_{NP} applied. The subscript μ has been introduced to define the new boundaries of the space charge region for the crystal.

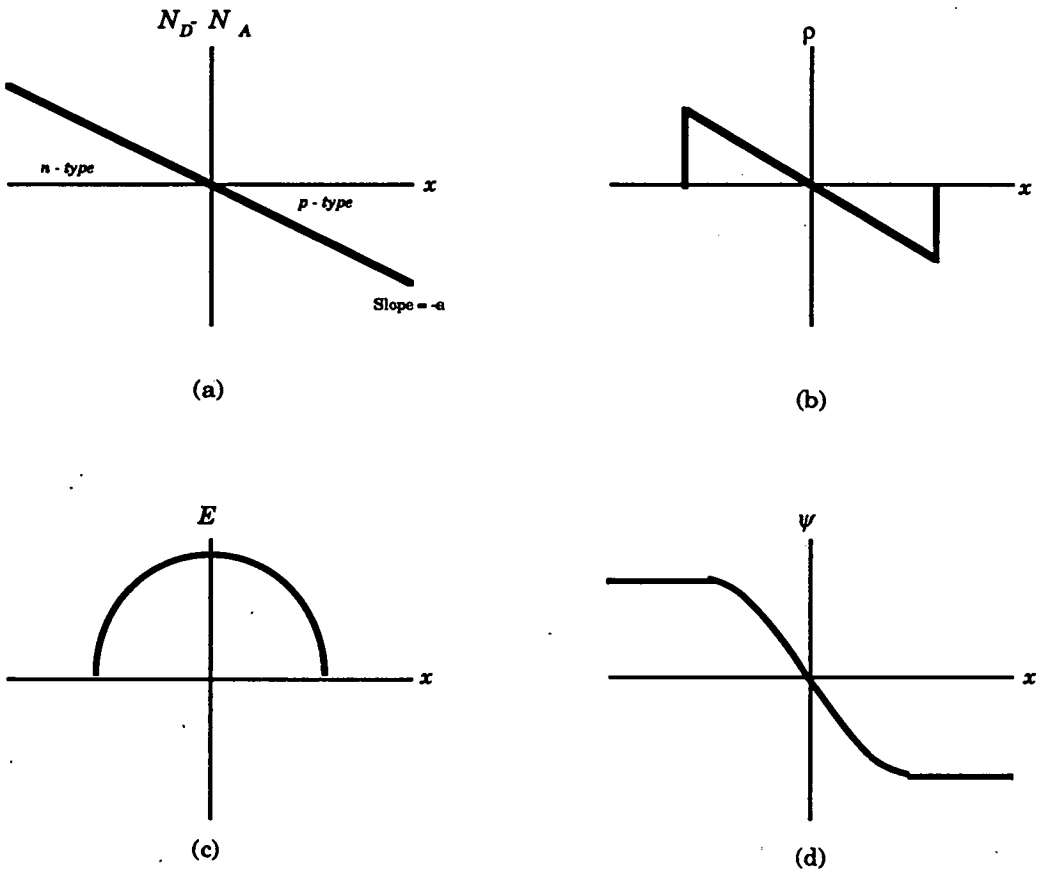


Figure 2.5 Properties of a linearly graded junction using the depletion approximation: (a) net dopant concentration: $N_D - N_A = -ax$, (b) space charge, (c) electric field, (d) potential.

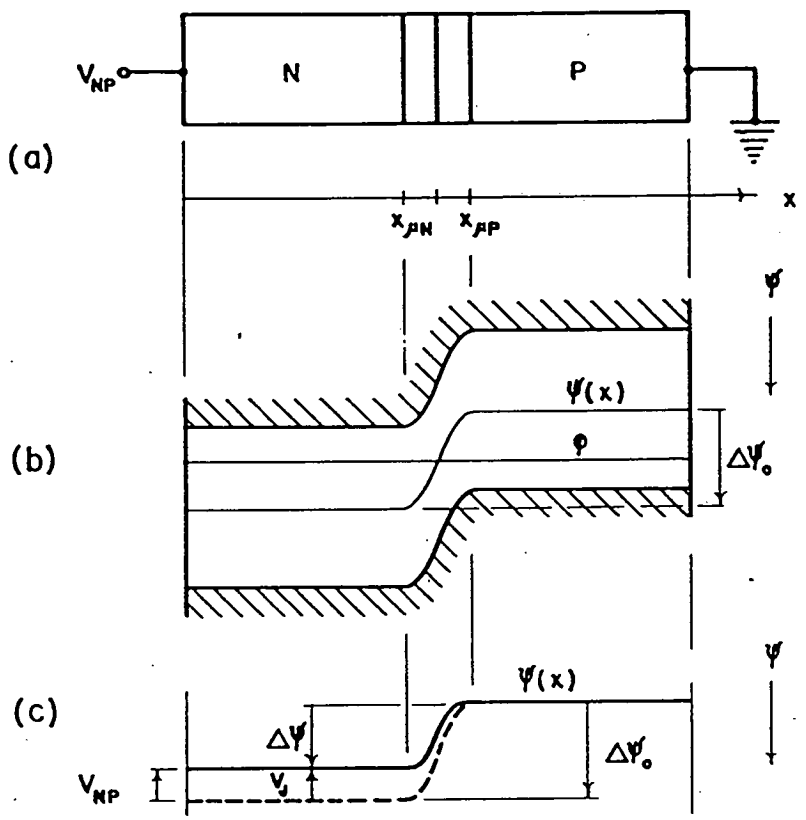


Figure 2.6 A symmetric step junction. (a) Physical representation with P-type side as a voltage reference. (b) Band diagram with no applied bias. (c) Potential profile for a small forward bias defining the imposed junction voltage V_j. Taken from reference [19].

From the application of Boltzman statistics to the semiconductor problem, the ratio of electron densities (or hole densities) is exponentially related to the potential difference between the two points when the sample is at equilibrium. Thus, invoking the Boltzman relation,

$$\frac{p_o(x_1)}{p_o(x_2)} = exp\left[U_o(x_2) - U_o(x_1)\right] \tag{2.16}$$

where $U_o(x)$ is the potential at point x_1 in the 1-D sample and $p_o(x_1)$ is the hole density.

$$\frac{p_{oN}}{p_{oP}} = exp\frac{q}{kT}\left[\psi(x_{\mu P}) - \psi(x_{\mu N})\right] \tag{2.17}$$

where p_{oN} and p_{oP} denote the neutral equilibrium values for holes on the p-side and the n-side. Solving this for minority carrier densities using

$$\psi(x_{\mu P}) - \psi(x_{\mu N}) = -\Delta\psi_o \tag{2.18}$$

gives

$$p_{oN} = p_{oP}\, exp\left[\frac{-q\Delta\psi_o}{kT}\right]. \tag{2.19}$$

Consider now the application of a small negative terminal voltage V_{NP}, shown in figure 2.6a. The low-level assumption lets us conclude that all the potential is dropped across the junction and not at the ohmic end contacts. As in figure 2.6c the change in barrier height is:

$$V_j = \Delta\psi - \Delta\psi_o \tag{2.20}$$

This is the imposed junction difference which is equal to the terminal potential in the low level example given here. Under high level conditions, there are a number of important differences. Some of these are discussed in section 2.2.2.

To obtain the desired boundary values of minority densities, $p_{\mu N}$ holds for low level injection as well as equilibrium. Then

$$p(x_{\mu N}) = p(x_{\mu P}) \cdot exp\frac{-q\Delta\psi}{kT}. \qquad (2.21)$$

Expression 2.21 is known as the Boltzman quasi-equilibrium value and can be can be re-written using 2.20 as

$$p(x_{\mu N}) = p(x_{\mu P})\, exp\left(\frac{-q\Delta\psi_o}{kT}\right) exp\left(\frac{-qV_j}{kT}\right). \qquad (2.22)$$

The strict preservation of charge neutrality dictates that when minority carriers are injected then the majority carrier concentration must rise by the same amount. One must remember that this amount of injected carriers may change the minority carrier density by orders of magnitude while the same amount makes relatively little difference to the majority carrier density.

For low level conditions, by definition, the majority carrier increment is negligible. Thus

$$p(x_{\mu P}) \approx p_{oP} \qquad (2.23)$$

substituting in equation 2.19 then re-writing 2.22 gives

$$p(x_{\mu N}) = p_{oN}\, exp\left(\frac{-qV_j}{kT}\right) \approx p_{oN}\, exp\left(\frac{-qV_{NP}}{kT}\right). \qquad (2.24)$$

Parallel arguments result in

$$n(x_{\mu P}) = n_{oP}\, exp\left(\frac{-qV_j}{kT}\right) \approx n_{oP}\, exp\left(\frac{-qV_{NP}}{kT}\right) \qquad (2.25)$$

equations 2.24 and 2.25 are known jointly as the law of the junction. It was first derived by Shockely [1] and the pair of equations are often referred to as the Shockley boundary conditions. Shockley used the concept of Quasi-Fermi levels to develop (2.24) and (2.25).

2.1.5 Current Voltage characteristics of the Ideal Junction

The one-sided or grossly asymmetric pn junction exhibits asymmetric carrier injection, this makes it a useful example for analytical treatment. Also, a practical "emitter" in a BJT is made this

way because asymmetric injection is specifically desired (chapter 3). Such a junction is shown in figure 2.7.

If we forward bias the junction with a negative voltage V_{NP} to the n-side to produce the profiles shown with heavy lines. According to the law of the junction, equations 2.24 and 2.25, the minority densities at both boundaries will be increased by the factor $exp(-qV_{NP}/kT)$. Let the factor be 10^3 so that from equation 2.25

$$V_{NP} = \frac{-kT}{q} ln\ 10^3 = -0.173 \text{ Volt}. \tag{2.26}$$

This value is well in the low-level injection regime. The next assumption is of major importance for this analysis. That is, the hole diffusion current at the left of the boundary is negligible compared to the electron diffusion current at the right hand boundary. In short, current passing through the forward biased n^+p junction consists almost exclusively of electrons.

For clarity, the $n(x)$ profile has been redrawn using a linear scale in figure 2.7c. The x origin has been placed at the right boundary to simplify future equations. We assume that the p region is extensive enough so that $n'(x)$ can be treated as purely exponential. The current at $x=0$ is purely diffusive and hence because electron current greatly exceeds hole current at the left hand boundary, it follows that the total current can be written

$$J \approx J_n(0) = qD_n \left.\frac{dn}{dx}\right|_{x=0}. \tag{2.27}$$

Excess electron density n' is defined as the difference between the actual density and n_o or

$$n' = n - n_{op} \tag{2.28}$$

so that

$$\frac{dn'}{dx} = \frac{dn}{dx} \tag{2.29}$$

thus equation 2.27 can be re-written as

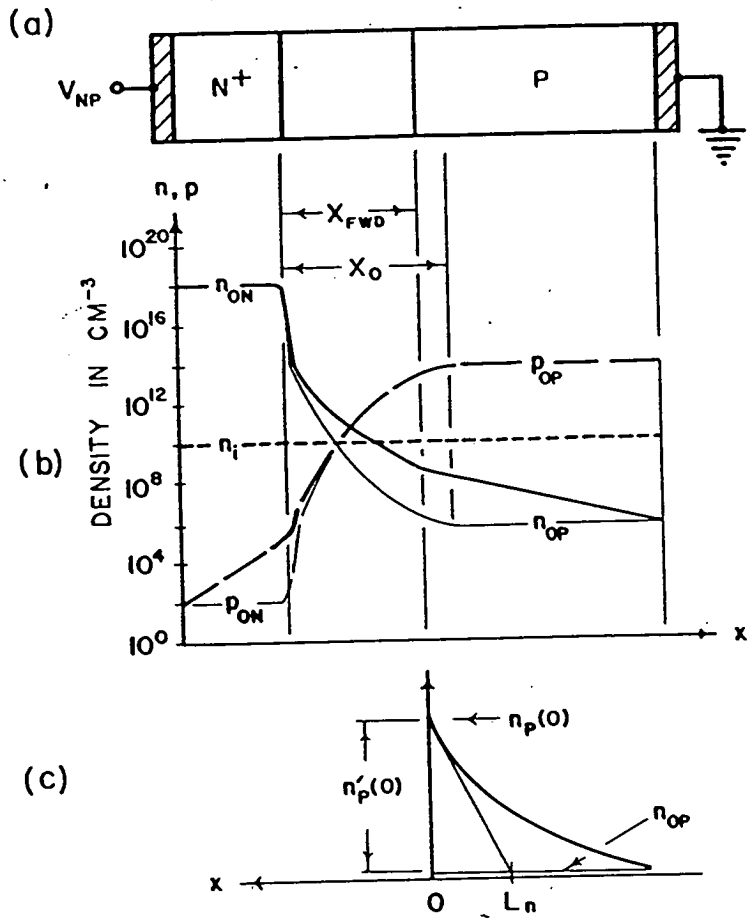


Figure 2.7 One sided (N+P) step junction under forward bias. (a) Physical representation. (b) Carrier profiles in log-linear representation. (c) Minority-electron profile in P region with fully linear presentation. Taken from reference [19].

$$J = qD_n \left. \frac{dn'}{dx} \right|_{x=0} . \tag{2.30}$$

We have assumed that $n'(x)$ is purely exponential thus

$$\left. \frac{dn'}{dx} \right|_{x=0} = -\frac{n'(o)}{L_n} \tag{2.31}$$

where L_n is the electron diffusion length. Using equation 2.28 again it follows that equation 2.30 becomes

$$J = -\frac{qD_n}{L_n} \left[n(o) - n_{op} \right] . \tag{2.32}$$

The law of the junction, equation 2.25, gives us $n(o)$, so that

$$-J = \frac{qD_n n_{op}}{L_n} \left[exp\left(\frac{-qV_{NP}}{kT} \right) - 1 \right] \tag{2.33}$$

all that remains to obtain a current voltage equation is to multiply both sides by junction area A, yielding

$$-I = \frac{qD_n n_{op} A}{L_n} \left[exp\left(\frac{-qV_{NP}}{kT} \right) - 1 \right] . \tag{2.34}$$

The coefficient in equation 2.34 is often termed the saturation current I_o, and that in equation 2.33 the saturation current density, J_o. This is because reverse bias causes the exponential term in either expression to "drop out"; thus the simple theory predicts that reverse current quickly levels out or "saturates" at I_o, or current density, at J_o.

$$-I = I_o \left[exp\left(\frac{-qV_{NP}}{kT} \right) - 1 \right] . \tag{2.35}$$

2.1.6 Generation and Recombination

The analysis of the pn junction that led to the diode equation was based on conditions in the quasi-neutral regions. The space charge layer was treated solely as a barrier to the diffusion of majority carriers, and it played a role only in the establishment of minority carrier densities at its boundaries. This is a reasonable first-order description of events and the

equation derived from it, equation 2.35, is called the ideal diode equation. Over a significant range of useful biases, however, the ideal diode equation is quite inaccurate, especially for silicon pn junctions. It is necessary to consider corrections to this equation that arise from a more complete treatment of events in the space-charge region. The main effect to be considered is carrier generation-recombination.

Many texts provide a good account of the Hall-Shockley-Read generation-recombination theory [6,7,8,9,10,11]. The inclusion of generation-recombination effects results in a modified current-voltage equation for the silicon diode.

$$-I = I_o\left[exp\left(\frac{-qV_{NP}}{kT}\right) - 1\right] + I_{o,gr}\left[exp\left(\frac{-qV_{NP}}{2kT}\right) - 1\right] \qquad (2.36)$$

The second term in equation 2.36 is a current which arises from generation-recombination in the space-charge region. This current will dominate the behavior of a silicon junction in the low-level part of its forward bias characteristics and almost all of it's reverse characteristic [12,13].

2.2 Principles of Transistor Operation

The Bipolar Junction Transistor (BJT) consists of two pn junctions in close proximity; it therefore has three regions. Figure 2.8a shows a schematic of a simple bipolar transistor with the three regions labeled the emitter, the collector and the base. Figure 2.8b shows a one-dimensional model of a n^+pn bipolar transistor which will be used to develop analytic and quantitative values for the currents flowing in this structure.

The first grown junction transistors actually possessed a structure like that shown in Figure 2.8b. Even in modern BJTs the central portion of the transistor approximates to a one-dimensional case, hence the analysis of the one dimensional problem is relevant.

2.2.1 1-D Analysis of npn Device

Figure 2.9 shows the band diagram for a npn transistor (symmetrical step junctions have been chosen for simplicity). In Figure

2.9(a) the transistor is at equilibrium with no external bias applied. From the pn junction theory developed we know external biases can forward bias or reverse bias a pn junction. Figure 2.9b shows the band diagram of a BJT under it's most commonly used mode of operation, in the common emitter configuration. Here the base-emitter junction is forward biased and the base-collector junction reverse biased.

Forward biasing the base-emitter junction results in large quantities of electrons being injected from the emitter to the base. Most of these electrons diffuse across the base to reach the base-collector junction. The base-collector junction is reverse biased, thus the electrons which reach this junction are swept by the electric field into the collector. It should be noted that even though the base-collector junction is reverse biased, a current approximately the size of the injected emitter current flows through it. This is the principle of transistor action: a large current flows in a reverse biased junction due to the existence of a forward biased junction in its proximity [14].

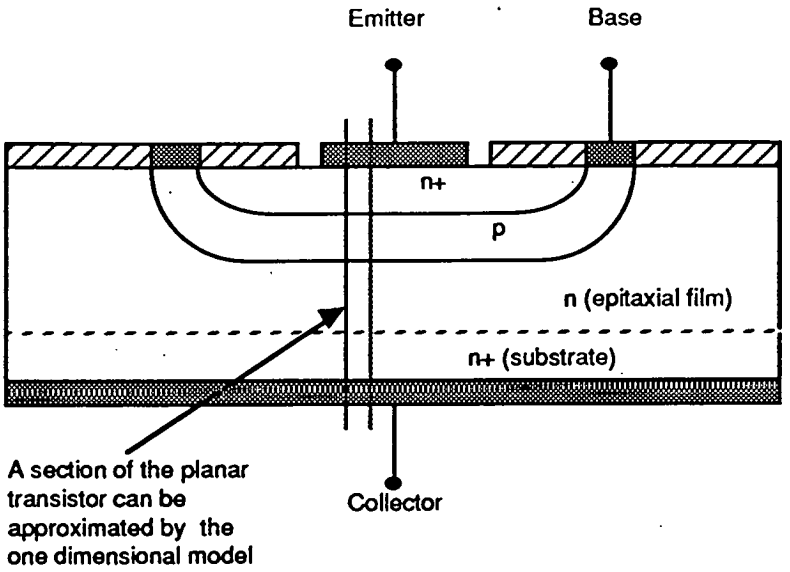
As stated above, not all electrons injected by the emitter will reach the collector, some will recombine with holes en-route through the base and the two space charge regions. Although the currents react in a complicated way they can be simply represented by Figure 2.10 [15].

The current shown by I_D is the defect current; if it did not exist the emitter would be "perfect". It is composed of holes which diffuse across the forward biased junction to recombine with electrons in the emitter.

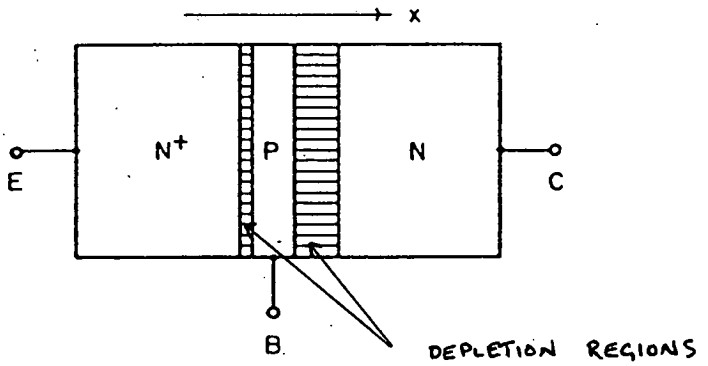
I_L is the leakage current, attributable to carrier generation within the base-collector space charge region. The holes which leak into the base can join the holes already present and contribute to the emitter defect current or they can recombine with electrons injected into the base. In modern transistors, however, recombination currents are usually neglected because they are so small.

The above assumption simplifies Figure 2.10 to give Figure 2.11. Thus I_B (the base current) becomes equal to the defect current I_D. Using theory expounded in the pn junction section above, we can develop analytic expressions for I_B and I_C using expressions for carrier transport across a junction with known boundaries. For a npn transistor with cross sectional area A:

$$I_C = AqD_{nB}\frac{dn_B}{dx} \tag{2.37}$$



(a)



(b)

Figure 2.8 Representation of the npn bipolar junction transistor (BJT) (a) Schematic of grown junction type transistor. (b) Rudimentary geometrical form.

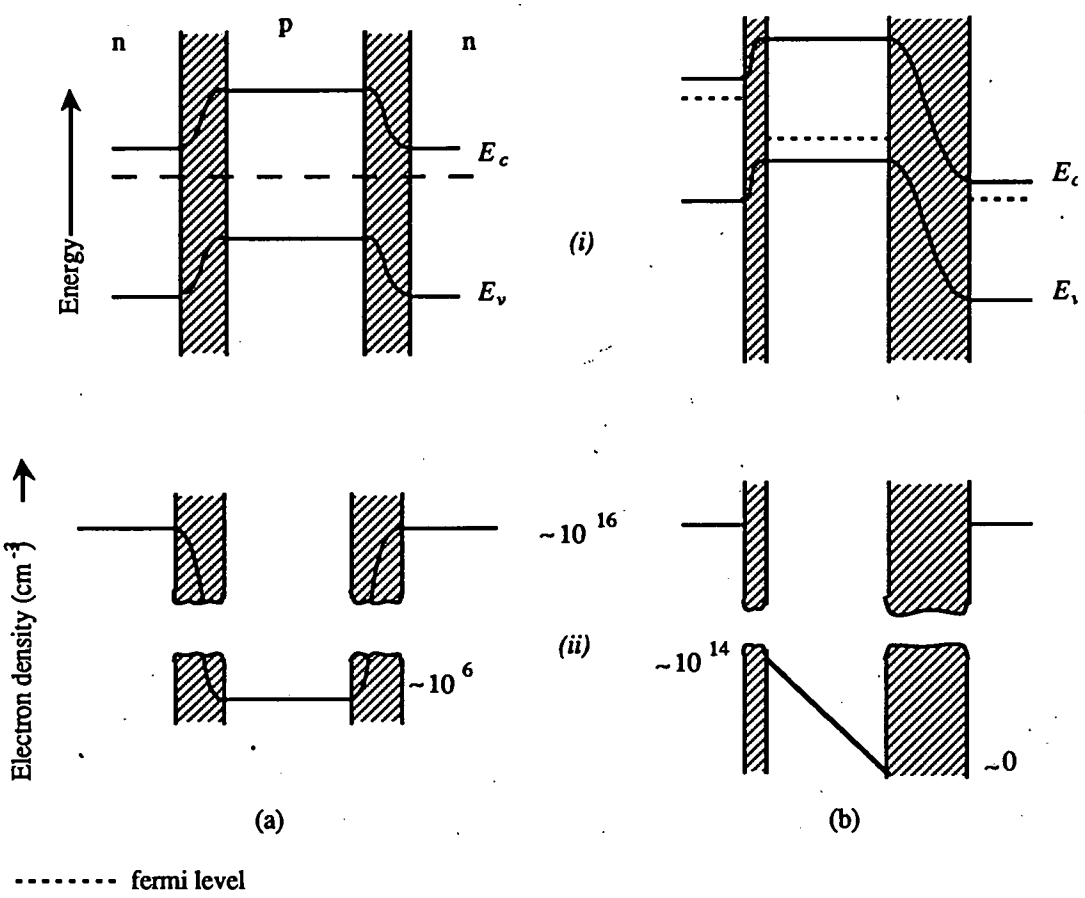


Figure 2.9 Energy-band diagrams (i) and electron-density sketches (ii) for the transistor sketched in figure 2.8. (a) Equilibrium condition, (b) one junction reverse-biased, and one junction forward-biased. The cross-hatched areas represent space-charge regions.

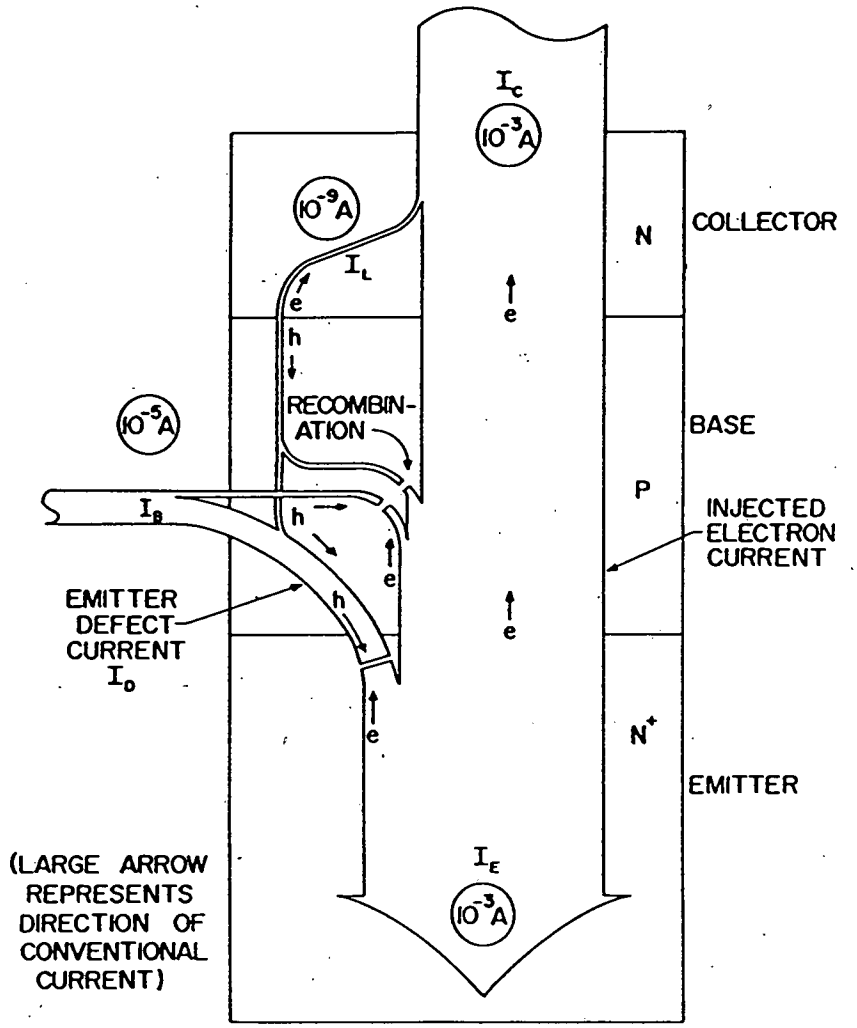


Figure 2.10 Various currents within the NPN BJT. Arrow widths represent current magnitudes qualitatively, and circled values indicate typical magnitudes in a small device. Breaks in arrows represent the condition of holes meeting electrons to recombine. Taken from reference [15].

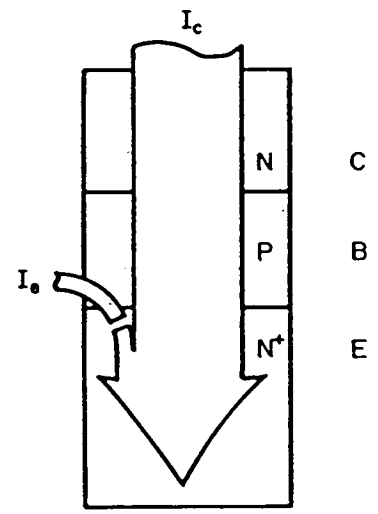


Figure 2.11 Simplified BJT current patterns showing only the two most important components. Note that these approximate I_B and I_C respectively.

For a positive terminal current we are interested in the magnitude of I_C.

$$I_C = \left| AqD_{nB}\left(\frac{-n_B(0)}{X_B}\right)\right| \tag{2.38}$$

Then using the law of the junction

$$I_C = \frac{AqD_{nB}n_{oB}\,exp\left(\frac{qV_{BE}}{kT}\right)}{X_B}. \tag{2.39}$$

Using the theory developed for the forward biased junction, the emitter defect current can be written as

$$I_D = \left| -AqD_{pE}\frac{dp_E}{dx}\right| \tag{2.40}$$

evaluated at the boundary of the emitter junction space charge layer. But the magnitude of this current is simply I_B, so that

$$I_B = |I_D| \approx \frac{AqD_{pE}p_{oE}\,exp\left(\frac{qV_{BE}}{kT}\right)}{L_{pE}}. \tag{2.41}$$

In the common emitter configuration I_B is the input current and a DC current gain can be defined as

$$\beta \equiv \frac{I_C}{I_B} \equiv h_{FE}\ . \tag{2.42}$$

Dividing equation [2.39] by [2.41] we obtain

$$\beta = \frac{D_{nB}n_{oB}/X_B}{D_{pE}p_{oE}/L_{pE}}. \tag{2.43}$$

From [2.43] it can be seen that β is fixed by constants defined by the BJTs structure. This is very important. For a large β, a small value of X_B is required and also a large ratio of n_{oB}/n_{op} which means a large ratio of N_D in the emitter to N_A in the base. The remaining quantities D_{nB} and D_{pE} are not under the control of the transistor designer to the same

degree as the others. β is a fundamental parameter in the Ebers Moll equations which are discussed in chapter 3.

2.2.2 Non Ideal Factors

The Early effect as it has come to be known was first analyzed by J.M. Early [16] in the era of grown junction BJTs. Although the effect was less pronounced in these devices, he noted a decrease in base width with increasing collector voltage. The effect is particularly pronounced in the alloy-junction BJT, and figure 2.12 uses this structure to explain the effect. Figure 2.12a shows the 1-D structure and figure 2.12b shows the base region profile with a small applied V_{CB}. When V_{CB} is increased the base-collector space charge region extends almost exclusively into the neutral base region. The base width X_B decreases and $\frac{dp}{dx}$ in the base region increases accordingly. This leads to an increase in terminal current I_C. Therefor, β has become voltage dependent. The base current does not increase significantly since it is primarily due to effects near the emitter-base junction [17].

Figure 2.12d demonstrates the effect on the output characteristics of the device. Extrapolation of the slopes of the output curves leftwards, will for many transistors converge at a point on the voltage axis. This point is named the Early voltage [18]. In terms of V_A (the forward early voltage), the smaller the value the more variable is the transistor in terms of dynamic conductance and dynamic output resistance.

2.3 MOS Device Operation

A rudimentary analysis of the MOS device is given below. From this current-voltage equations will be derived. It will be shown how these form the basis for the level 1, 2 and 3 SPICE models discussed in chapter 3.

Figure 2.13 shows a schematic diagram of the MOSFET device showing its principal features. The MOSFET consists of a semiconductor region, a thin insulator (silicon dioxide) overlying it and a conductive material forming an upper plate for the capacitor (the gate). The n^+ regions in the p-type semiconductor are termed source and drain because electrons (in this case) can flow from source to drain

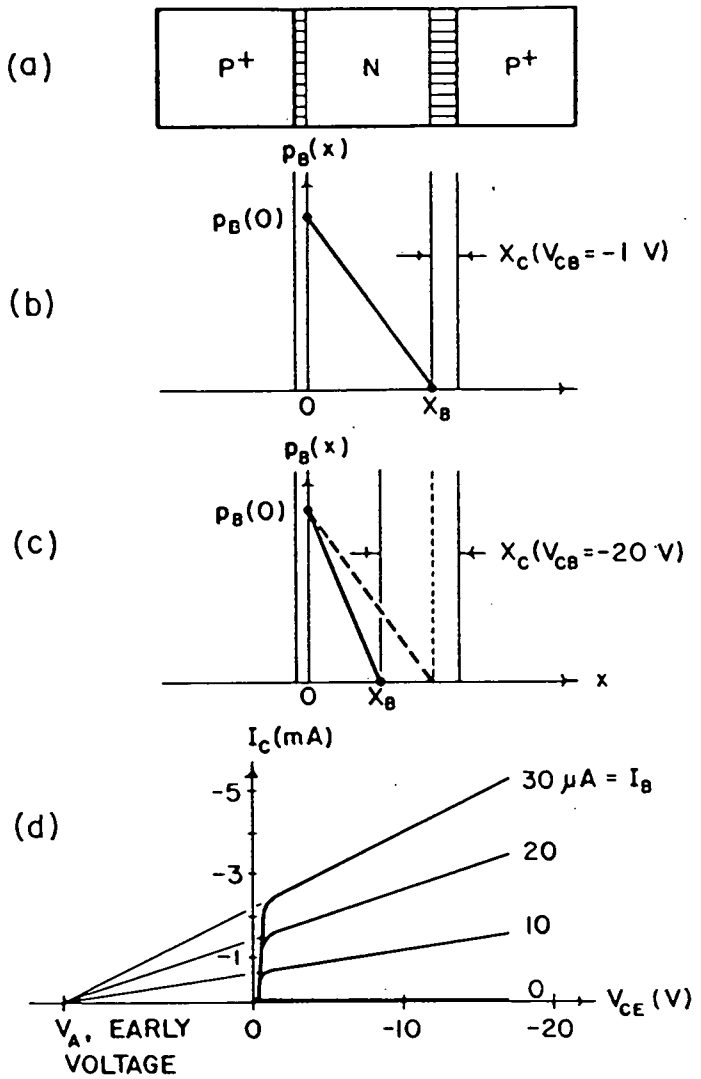


Figure 2.12 The Early effect. (a) Physical representation of a BJT exhibiting the Early effect. (b) Base-region minority hole profile with a small value of V_{CB} . (c) Base-region minority-hole profile with a much larger value of V_{CB} . (d) Output plane, showing V_{CB} -modulated collector current I_C resulting from changes in active base thickness X_B caused by collector-depletion-layer encroachment. Taken from reference [19]

when the silicon region near the oxide-silicon interface permits electron conduction. This occurs when a gate voltage is applied sufficient to cause inversion. Biasing the gate positively with respect to the silicon attracts electrons to the interface region. When their volume density (which peaks immediately below the oxide) equals the equilibrium hole density in the underlying p-type silicon, the device is said to be at the threshold of strong inversion. The gate voltage corresponding to this condition is termed the threshold voltage V_T. Once the gate has been biased at a voltage in excess of V_T, a positive voltage applied to the drain region with respect to the source causes electron transport from source to drain through the inversion layer. This constitutes the channel. Since electrons populate the channel in the example above, this device is an n-channel MOSFET.

Figure 2.14 shows the structure of figure 2.13 in cross-section and defines a set of bias conditions for further analysis. To start, let the source and drain be electrically common as shown in 2.14. It is customary to refer all bias voltages to the source. Thus in the case of the gate, the potential of interest is

$$V_G - V_S \equiv V_{GS} \ . \tag{2.44}$$

Let us assume a value $V_{GS} > V_T$, so that an inversion layer exists. To cause electrons to flow from source to drain let us also apply a voltage to the drain,

$$V_D - V_S \equiv V_{DS} \tag{2.45}$$

and let $V_{DS} << V_{GS}$. Hence we have established conditions for a channel current I_{ch}, where the conventional current flows from drain to source. The condition $V_{DS} << V_{GS}$ causes the channel to have a nearly constant "thickness" from source to drain. Using the concept of sheet resistance R_S we may write the resistance of a rectangular layer as

$$R = R_S \frac{L}{Z} \tag{2.46}$$

where L is the layer dimension in the current direction, and Z is that orthogonal to L, as illustrated in figure 2.15. Given sheet resistance no

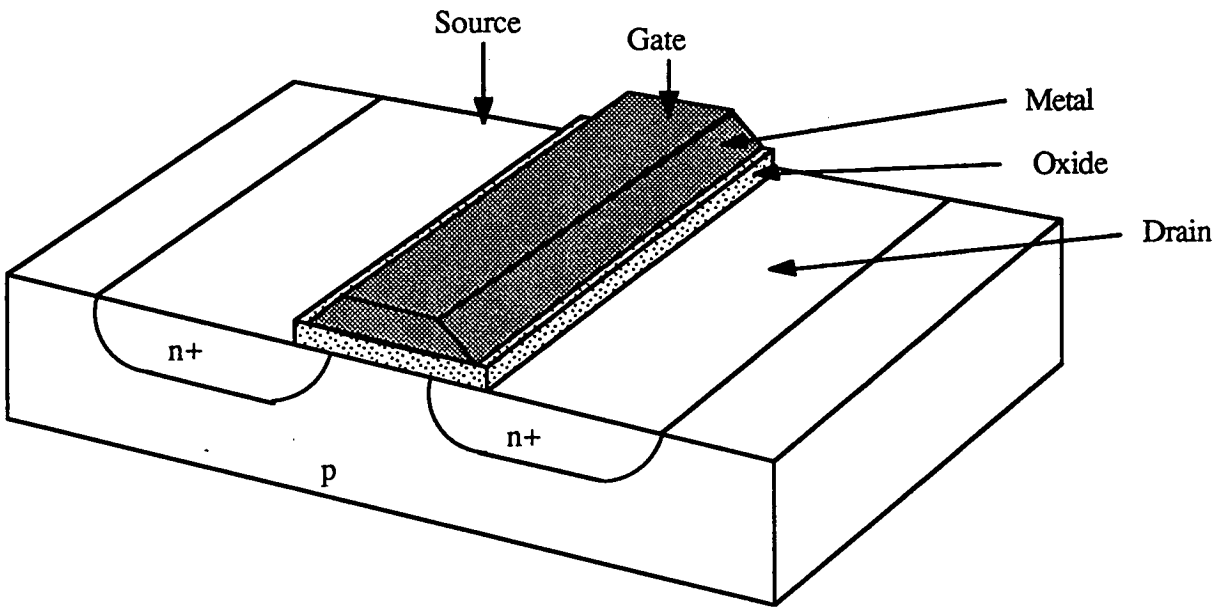


Figure 2.13 Schematic diagram of a MOSFET, showing principal features.

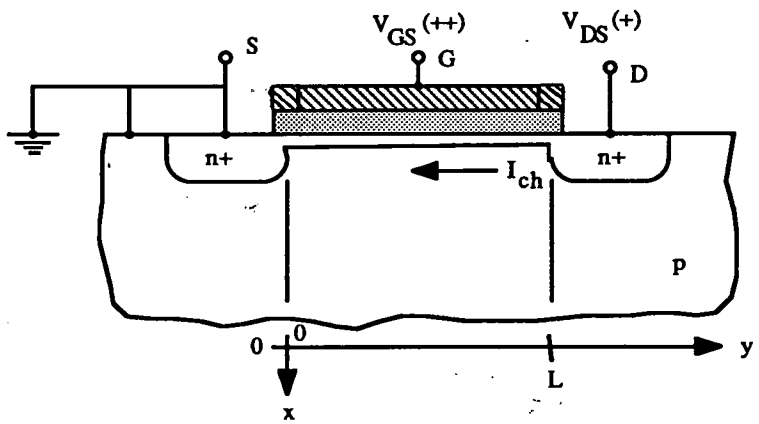


Figure 2.14 Cross-sectional diagram of n-channel MOSFET with grounded source and substrate terminals, with gate voltage V_{GS} sufficient to produce an inversion layer (i.e., to form a channel), and with a smaller drain voltage V_{DS}.

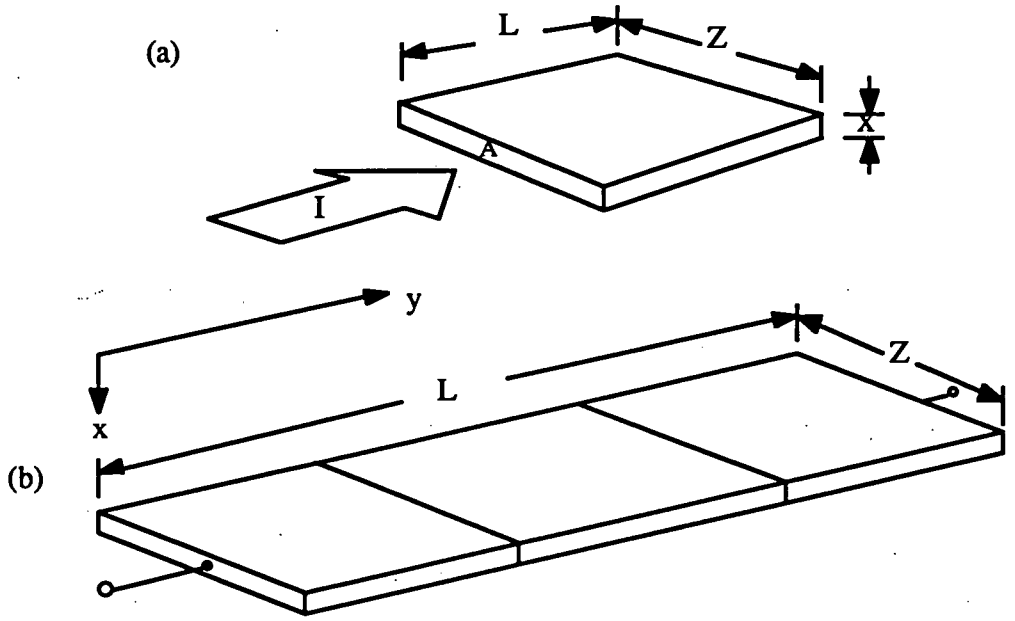


Figure 2.15 Formulating the sheet-resistance concept. (a) A "square" of conducting material having a thickness X and a cross sectional area A = LX. (b) When squares are assembled in series, the resulting resistance is found by multiplying the sheet resistance R_s by the aspect ratio of the assembly, L/Z.

knowledge is required of the samples thickness nor resistivity. Uniformity in the x direction is not required. A channel is an example of a thin sheet that is non-uniform in the x direction.

The ratio L/Z specifies the number of squares in series; the resistance of each times that ratio gives the total resistance. When $L < Z$, the ratio Z/L gives the number of squares in parallel. The net resistance is the resistance of each square divided by Z/L. If we assume that the channel has a well-defined thickness X, and is uniform and has a volume of density n of electrons, then the amount of charge per unit area in figure 2.15b can be written

$$Q_n = -qnX \tag{2.47}$$

using

$$Rs \equiv \frac{\rho}{X} \ . \tag{2.48}$$

We can eliminate X

$$Q_n = -qn\frac{\rho}{R_s} = -\frac{qn}{R_s}\frac{1}{q\mu_n n} = \frac{1}{\mu_n R_s} \ . \tag{2.49}$$

Next return to the case of the channel ,using the capacitor law

$$Q = CV \tag{2.50}$$

and the foregoing approximations we can write

$$Q_n = -C_{ox}(V_{GS} - V_T) \ . \tag{2.51}$$

The minus sign enters because V_{GS} is applied to the field plate, while Q_n resides in the other plate of the MOS capacitor. That is, the channel or inversion layer may be considered to commence at the voltage V_T. Since the electrons are so close to the oxide-silicon interface, the MOS capacitor at this point may be regarded as a simple parallel plate capacitor and it's charge per unit area is determined, via the capacitor law, by the applied voltage in excess of V_T. From equations 2.49 and 2.50

$$R_S = -\frac{1}{\mu_n Q_n} = \frac{1}{\mu_n C_{ox}(V_{GS} - V_T)} \tag{2.51}$$

for the MOSFET in figure 2.13. We also assume that there is actually a potential variation along the channel, owing to the channel's finite resistance and the passage of I_{ch}, though we have assumed this variation to be small. The effect of this potential drop is to cause a variation with y of channel thickness or charge per unit area. This is shown in an exaggerated fashion in figure 2.16. This is the gradual channel approximation. Because the polarities of V_{GS} and V_{DS} are the same, the effect of the $V(y)$ variation is to cause less voltage drop through the oxide near the drain end of the channel than near the source end. Consequently, using the gradual channel approximation, the channel grows thinner (has fewer electrons per unit area) toward the drain end. This means that $R_s = R_s(y)$, or, sheet resistance varies spatially in the channel. From equation 2.51

$$R_s(y) = \frac{1}{\mu_n C_{ox}[V_{GS} - V_T - V(y)]}. \tag{2.53}$$

Employing the sheet resistance equation 2.48, we can write for the element dy of channel length

$$dR = R_s \frac{dy}{Z} . \tag{2.54}$$

Where Z is the dimension of the channel in the Z direction. Combining equations 2.53 and 2.54 gives

$$dR = \frac{dy}{Z\mu_n C_{ox}\left[V_{GS} - V_T - V(y) \right]} . \tag{2.55}$$

From Ohm's law

$$I_{ch} = -\frac{dV}{dR} \tag{2.56}$$

where V will be used to mean $V(y)$. The negative sign enters because conventional current flows "down" a voltage gradient, while a positive dV means a voltage rise. Combining equations 2.55 and 2.56 yields

$$I_{ch} = -\frac{Z\mu_n C_{ox}\left[V_{GS} - V_T - V\right] dV}{dy}. \tag{2.57}$$

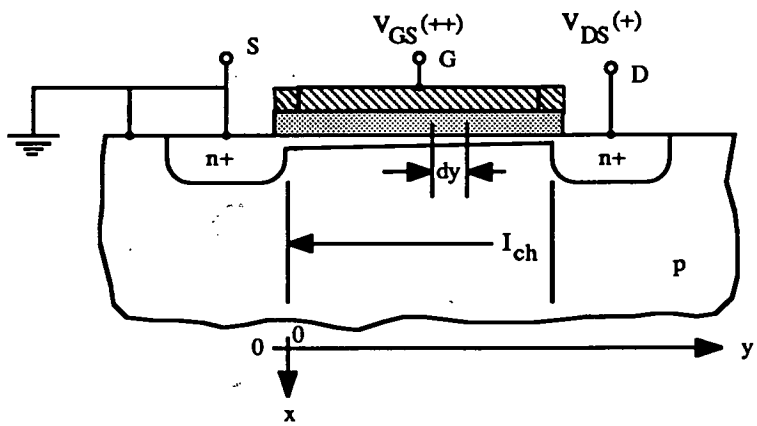


Figure 2.16 Cross-sectional diagram of a MOSFET for purposes of rudimentary analysis. Beyond-threshold conditions prevail throughout the channel.

Note that $I_{ch} \neq f(y)$. This is because carriers in the channel are confined to a potential well near the surface from which they cannot escape, and consequently, channel current must be constant from source to drain. Separating variables and integrating from source to drain gives us

$$I_{ch} \int_0^L dy = -Z\mu_n C_{ox} \int_0^{V_{DS}} (V_{GS} - V_T - V) \, dV \tag{2.58}$$

or

$$I_{ch} = -\frac{\mu_n C_{ox}}{2} \frac{Z}{L} \left[2 (V_{GS} - V_T) V_{DS} - V_{DS}^2 \right]. \tag{2.59}$$

Now $I_{ch} = -I_D$ where I_D designates a terminal current. The negative I_{ch} indicated in figure 2.16 flows leftward, which causes it to flow into the drain terminal. An inward-flowing terminal current is, however, conventionally designated as positive, which accounts for the sign difference. Consequently

$$I_D = \frac{\mu_n C_{ox}}{2} \frac{Z}{L} \left[2 (V_{GS} - V_T) V_{DS} - V_{DS}^2 \right]. \tag{2.60}$$

The oxide capacitance per unit area C_{ox} is

$$C_{ox} \equiv \frac{\varepsilon_{ox}}{X_{ox}} \tag{2.61}$$

where $\varepsilon_{ox} \equiv$ the absolute permativity of the oxide and $X_{ox} \equiv$ oxide thickness. Equation 2.60 constitutes the SPICE level 1 model for the MOSFET device. This is the starting point for the discussion of the model and it's refinement in chapter 3.

References

[1] J. M. Warner and B. L. Grung,"Transistors Fundamentals for the Integrated Circuit Engineer,"*Wiley-Interscience*,p. 294, 1983.

[2] W. Shockley,"The Theory of p-n junctions in Semiconductors and p-n junction transistors,"*Bell Sys. Tech. J.*, 28, pp. 435, 1949.

[3] C. T. Sah, R. N. Noyce and W. Shockley,"Carrier Generation and Recombination in p-n Junctions and p-n Junction Characteristics," *Proc. IRE.*, 45, pp. 1228, 1957.

[4] J. L. Moll,"The evolution and the Theory of the Current Voltage Characteristics of p-n Junctions,"*Proc. IRE. 46*, p. 1076, 1958

[5] W. Shockley,"The path to the conception of the Junction Transistor,"*IEEE. Trans, vol ED. 31, no 11*, p. 1523, 1984.

[6] R. N. Hall,"Electron-Hole Recombination in Germanium,"*Phys. Rev. 87*,p. 387, 1952.

[7] W. Shockley and W. T. Read, Jr.,"Statistics of the Recombination of Holes and Electrons,"*Phys. Rev. 87*, p. 835, 1952.

[8] S.M Sze,"Physics of Semiconductor Devices,"*Wiley 2nd edition*,p 90, 1981.

[9] A. S. Grove,"Physics and Technology of Semiconductor Devices,"*Wiley*, P129, 1967.

[10] R.M Warner and B. L. Grung,"Transistors,"*Wiley-Interscience*, pp. 265-283, 1983.

[11] R. S. Muller and T. I . Kamins,"Device Electronics for Integrated Circuits,"*Wiley 2nd edition*, pp224-226, 1986.

[12] A. S. Grove,"Physics and Technology of Semiconductor Devices,"*Wiley*, pp. 173-189, 1967.

[13] R.M Warner and B. L. Grung,"Transistors,"*Wiley-Interscience*, pp. 428-432, 1983.

[14] A. S. Grove,"Physics and Technology of Semiconductor Devices,"*Wiley*, p. 211, 1967.

[15] R.M Warner and B. L. Grung,"Transistors,"*Wiley-Interscience*, p. 491, 1983.

[16] J.M. Early," Effects of Space Charge Layer Widening in Junction Transistors,"*Proc. Ire. Vol.32*, pp. 1401-1406, 1952.

[17] A. S. Grove,"Physics and Technology of Semiconductor Devices,"*Wiley*, pp. 226-227, 1967.

[18] H.K. Gummel, H.C. Poon,"An Integral Charge Control Model of Bipolar Transistors," *Bell sys. Tech. J.*, Vol. 49, p. 827, 1970

[19] R.M Warner and B. L. Grung,"Transistors,"*Wiley-Interscience*, 1983.

Chapter 3

Bipolar and MOS Modelling

Introduction

This chapter will discuss process control and its role in semiconductor fabrication past and present. From these observations an attempt to forecast and classify the future role of process control will be made.

The origins of process control lie with the development of test vehicles to enable reliability studies to be made on MOS and bipolar LSI chips [1, 2]. Process control has now reached the point where parametric test details obtained from process control chips (PCCs) are used on a go-nogo basis for further functional testing, eg for catastrophic process errors such as missing implant or wrong mask. More commonly, parametric results are used to correct drifting processes or to pinpoint out of control equipment in a process step where more than one piece of equipment is available for a specific operation. The accumulation and analysis of this parametric data has now become a subject in itself with the advent of computer aided design/computer aided manufacture CAD/CAM software packages such as PROMIS [3] and COMETS. Their purpose is to monitor batch progress through the process and to keep records of both electrical and physical characteristics (ie which implanter or oxidation tube was used for a particular batch).

The field of parametric testing includes test structures, test chip architectures, relating parametric test data to processing and to functional test results, and parametric testers themselves [4]. This chapter will attempt to examine each of these aspects in order to put in

context the work done on this topic.

3.1 Parametric Test Systems

Before examining the parametric data available and methods of extracting it, some time should be devoted to looking at the equipment available for parametric extraction.

The first commercially available systems dedicated to parametric test arrived in the marketplace around 1975. These first generation PT systems had single task operating systems based on 8 or 16 bit processors such as the Z80 or DEC 11/03. Often these systems had only floppy disks for mass storage and collected parametric test data at rates of 1 to 10 results per second. These machines were used to automate manual curve tracer measurements [5]. Three major parametric systems emerged, the Accutest system 3000, Keithley system 300 and the Lomac LM-80. These testers competed favourably with semiconductor manufacturer's in-house test equipment by furnishing higher accuracies, faster tester speeds, fewer errors and real time data analyses [6, 7].

Whilst some of these systems are still in operation in mass production semiconductor fabrication, they have generally have been superseded by the new breed of parametric test systems which are more complex, faster and included as part of a computer integrated manufacturing package (CIM).

3.1.1 Tester uses

A parametric test system operated in a suitable environment can provide process monitoring information and electrical characterisation. In the case of process monitoring, tight control over such parameters as

(i) Sheet resistivities

(ii) Junction depths

(iii) Oxide thickness

(iv) Critical dimensions

can be obtained to give quick feedback to the product engineer for efficient front-end quality assurance. Several test chips per wafer are usually tested for this purpose. Parametric testing for process control must therefore be fast and efficient with the additional constraint that spurious data must be easily identified, i.e., data needing much analysis to determine whether it comes from a good device or a bad device is not appropriate for process control parametric test.

However, when parametric testing is used for electrical characterisation, some tests require much analysis for their interpretation. Detailed CV techniques for impurity profile definition is one such example. Electrical characterisation may include the same tests as those in process control but would include more data points. A full characterisation of a test chip would be carried out if the routine parametric test in process control highlighted a deviation from the set tolerances in the process. The data obtained can then be correlated with the wafer's process history to obtain the reason for the failure.

The third category of use for a parametric test system is device characterisation. This would be done on an off-line parametric test system. It is here that the process engineer, device engineer and design engineer must collaborate. Devices fabricated by a new or altered process must be characterised to provide parameters for device and circuit models such as SPICE, or more frequently for large manufacturers, their own in-house models or variations of SPICE. In order to create a new design or modify an old one, the designer must have access to a best - worst case library of model parameters. Parametric testing is the cornerstone in the iterative cycle between design requirement and process possibilities. The need for faster and more frequent cycles in this relationship has caused process monitoring and device modelling to be two of the fastest growing applications of parametric testing.

3.2 Parametric testing for Statistical Process Control (SPC)

In order to identify the role of parametric testing for process control in a large semiconductor manufacturing organisation, a resume of Motorola's SPC guide is presented here. This highlights the company commitment to SPC. Motorola's expressed objective is the

achievement of "error free performance" in products and services. Whether zero defects is a consistently achievable goal or not remains subject to conjecture. However, Motorola is committed to a hundred-fold performance improvement by 1991, and a Six Sigma Capability by 1992.

3.2.1 Six Sigma Capability

Each process attempts to reproduce its characteristics identically from unit to unit. Inherent in each process, however, there are variations in conditions and materials that are uncontrollable and unalterable. In all cases, therefore, the unit-to-unit output characteristics may vary somewhat from the ideal (design target).

The performance of a product is determined by how much margin exists between the design specifications and the actual value of that specification. For some processes, such as those using real-time feedback to control the output, the variations can be quite small; for others they may be quite large. Many of the parametric data of a given specification tend to follow the normal distribution curve shown in Figure 3.1.

Variation of the process is measured in Standard Deviations (sigma) from the Mean. The normal deviation, defined as process width, is ±3 Sigma about the mean, representing a yield of 99.73%. The Motorola goal is to design their products with a component yield that is significantly better than ±3 Sigma. The table in Figure 3.1 shows that a design which can accept twice the normal ±3 Sigma variation of the process (design width = ±6 Sigma) will have a product yield of 99.9999998%, corresponding to 2 defective parts per billion. SPC will be used to achieve Motorola's goal of ±6 Sigma capability in product design and manufacturing by 1992.

3.2.2 Statistical Process Control

Process variations can result from two causes: Common Causes and Special Causes. Deviations resulting from Common Causes are those that are inherent in the process and cannot be reduced without changing the process itself. Their effects are reflected in the location, spread and shape of the distribution curve. Indeed, it is the Common Causes, alone, that determine the capability of a process, and any basic

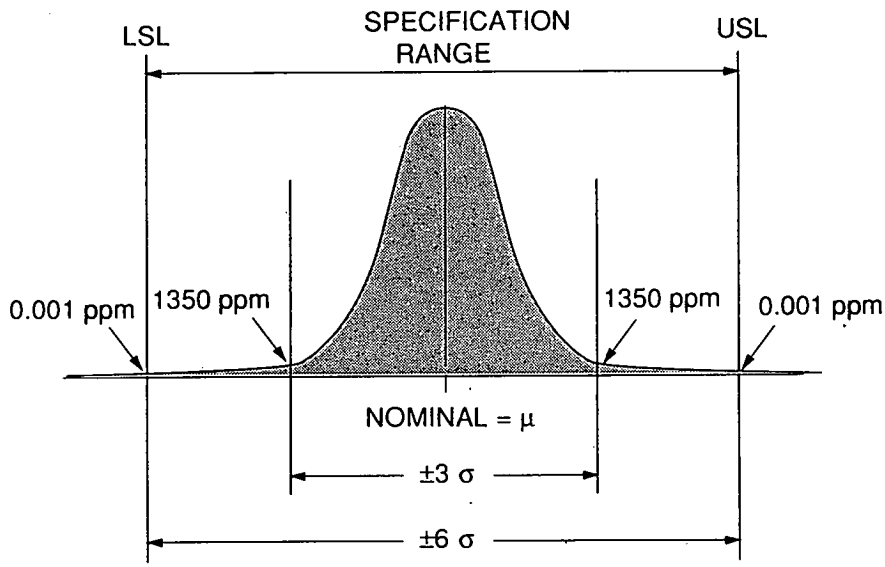


Figure 3.1 Typical normal distribution.

improvement in process capability can be achieved only be reducing or eliminating the deleterious effects of Common Causes.

Deviations due to Special Causes result from changes in external conditions that are controllable. Operator training and experience, changes in temperature or humidity, variations in test equipment are examples of Special Causes. A process can be considered to be under Statistical Process Control only when such Special Causes have been eliminated or adequately reduced. Their detection and elimination are the primary functions of statistical process control procedures. The work presented in this thesis demonstrates some novel test structures designed to help monitor the Special Causes present in a CMOS process.

3.3 Bipolar modelling

As discussed above the statistical control of a process can only be achieved through the detection and elimination of changes due to controllable external conditions. Device modelling allows the changes that are measured physically by parametric extraction, to be related to the internal structure of the device.

The models presented below will be referenced in later chapters. They are presented here in their simplest form with reference to more detailed derivations. The equations derived will be used later to support conclusions drawn about the relationships between parasitic devices and CMOS devices.

3.3.1 The Ebers and Moll equations

The equations to be derived here were first presented in 1954 by Ebers and Moll [8]. More than 35 years later some of the most complex computer modelling software still uses these equations as their basic framework. In this light the derivation presented here varies from that presented in their original paper. In this section we will derive the transport equations, ie the transported current across the junction boundaries. This is the form in which most computer simulation programs adopt the model and as such is most relevant for later discussion.

The Ebers and Moll model explains the operation of the BJT in terms of diode interaction (chapter 2.1). The transistor to be considered

here is an ideal 1-D npn BJT. The npn transistor was chosen since it is the most common of the BJTs and was treated in chapter 2. If we consider the common base configured transistor shown in figure 3.2, we can write

$$I_C = qAD_n \left|\frac{dn}{dx}\right| = qAD_n\left(\frac{n_B(0) - n_B(X_B)}{X_B}\right). \tag{3.1}$$

One of the explicit assumptions made by the Ebers Moll model was that the emitter and collector junctions were individually governed by current voltage equations of the form

$$-I = I_o\left[exp\left(\frac{q(-V_{NP})}{kT}\right)-1\right]. \tag{3.2}$$

We have seen earlier in chapter 2 equation 2.35 that this is the law of the junction first proposed by Shockley. By utilising the law of the junction, expressions for the minority carrier densities at the junction boundaries can be found :

$$n_B(0) = n_{0B}\, exp\left(\frac{q(-V_{EB})}{kT}\right) \tag{3.3}$$

and

$$n_B(X_B) = n_{0B}\ . \tag{3.4}$$

Using the boundary conditions above and equation 3.1 we have :

$$I_C = qAD_n n_{0B}\left(\frac{exp\left[\frac{q(-V_{EB})}{kT}\right]-1}{X_B}\right). \tag{3.5}$$

To follow convention let us say :

$$I_i \equiv \frac{qAD_n n_{0B}}{X_B} \tag{3.6}$$

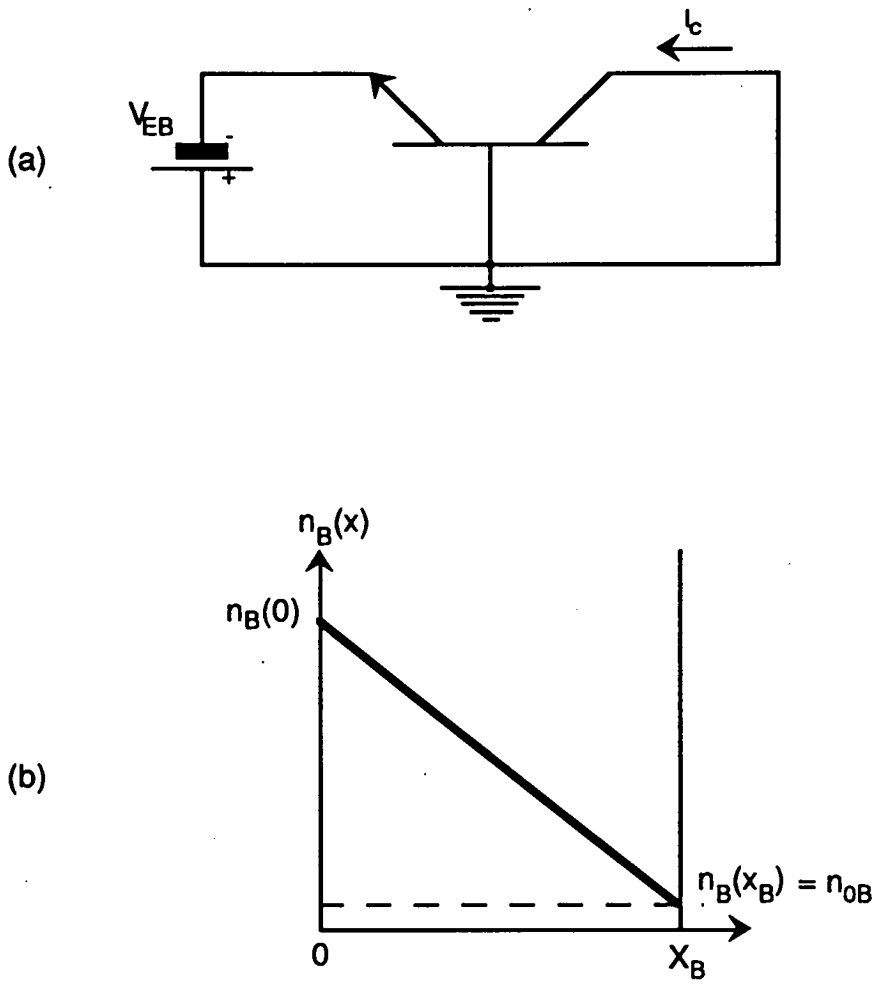


Figure 3.2 Forward common-base operation. (a) Bias configuration with emitter junction forward-biased and collector junction short-circuited. (b) Base-region minority-electron profile corresponding to (a) conditions. Law of the junction specifies equilibrium density at $x = X_B$.

then

$$I_C = I_i\left(exp\left[\frac{-q(V_{EB})}{kT}\right]-1\right). \tag{3.7}$$

At this point the physical significance of I_i should be noted. It has the form of a diffusion current whose gradient is n_{0B}/X_B. This current represents a parameter which is dependent on transistor structure and can be obtained in principle by short circuiting the emitter junction and reverse biasing the collector junction. This arrangement is shown in figure 3.3a.

The reverse bias voltage V_{CB} can range from 0.2V up to large values without altering the gradient of the carrier profile. This is illustrated in figure 3.3b which shows the minority carrier profile for two different bias voltages. This condition holds true providing junction boundary motion is neglected. This was another explicit assumption made by Ebers and Moll. Note that a coordinated change in X_B and equilibrium minority carriers leaves I_i unaltered. That is I_i is fixed if the gradient is fixed.

$$\frac{n_{0B}}{X_B} = constant\,. \tag{3.8}$$

Invoking the mass action law :

$$\frac{n_i^2}{N_{AB}X_B} \cong \frac{n_{0B}}{X_B} = constant\,. \tag{3.9}$$

Thus I_i is fixed by the quantity $N_{AB}X_B$ which has dimensions of number per unit area and gives the areal net impurity of the active base region.The quantity $N_{AB}X_B$ is known as the Gummel number of the transistor. This parameter gives important information about the static properties of the BJT [9]. What we have derived so far is a parameter I_i that characterises the entire base region rather than a junction region. Direct measurement of I_i is not straightforward. However, the indirect method of measurement introduced by Gummel is. The method is illustrated in figure 3.4.

By choosing bias values large enough so that equation 3.2 can be regarded as purely exponential, but small enough so that low level

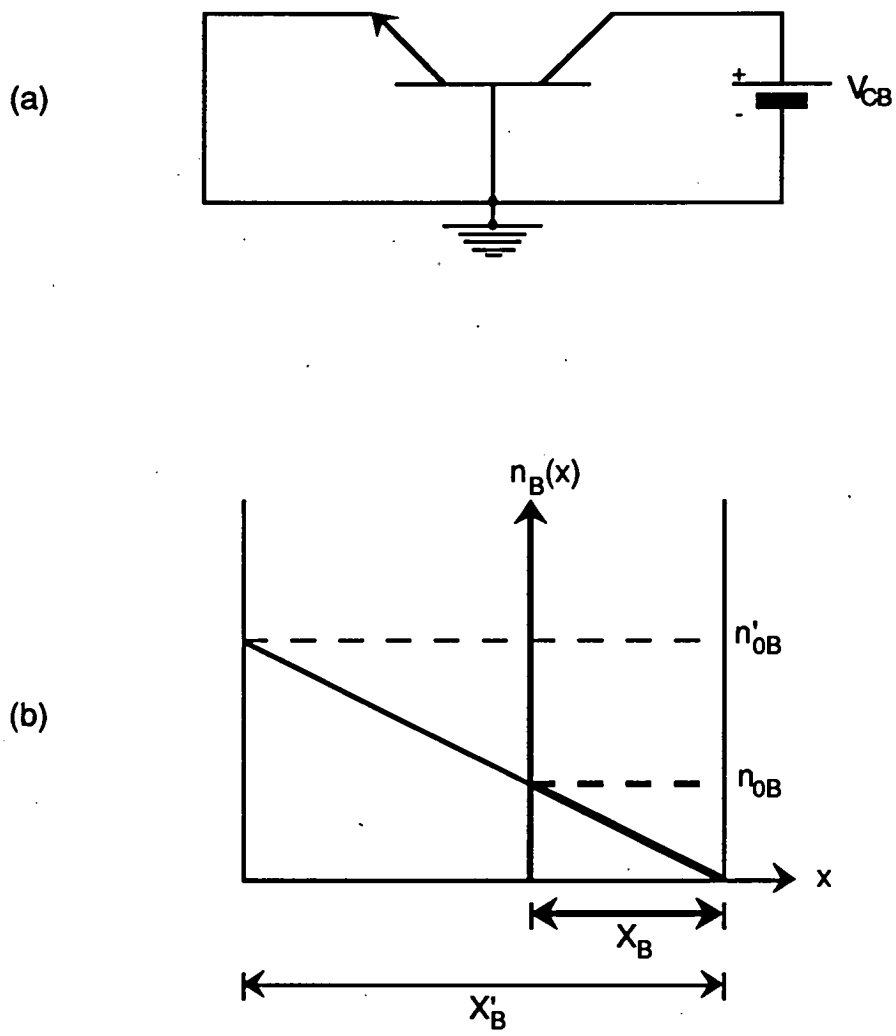
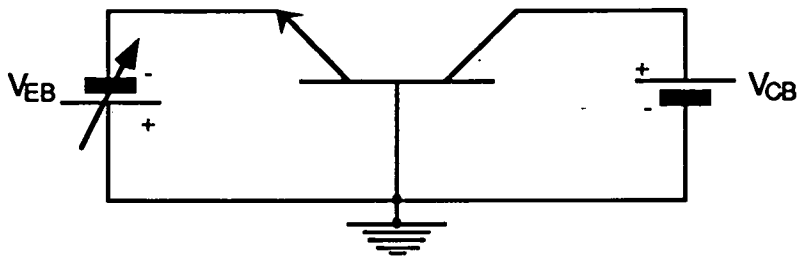


Figure 3.3 Defining the intercept current I_i . (a) Circuit configuration with emitter junction short-circuited and collector junction reverse-biased. (b) Corresponding electron profile in the base region (heavy lines). Coordinated change of X_B and n_{0B} leaves dn/dx (and hence I_i) unaltered.

(a)



(b)

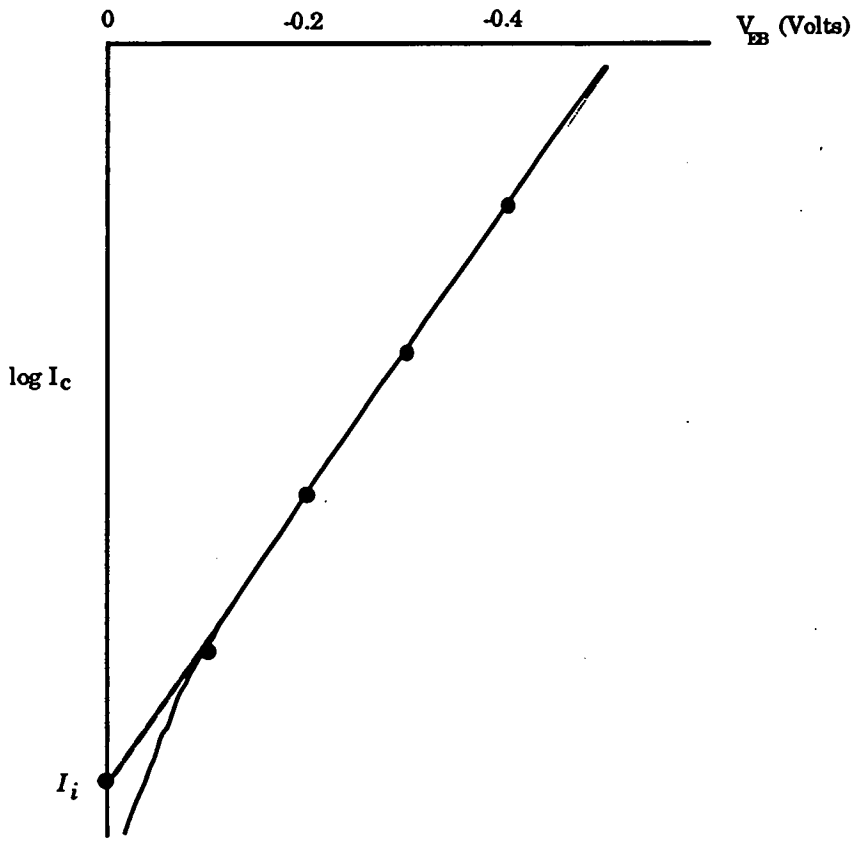


Figure 3.4 Determination of intercept current I_i . (a) Bias configuration with variable V_{EB}. (b) Plot log I_C vs. V_{EB} and extrapolate to $V_{EB} = 0$, to determine a current corresponding to a profile analogous to that in figure 3.3b.

conditions still exist, one can extrapolate the log $|I_C|$ vs $(-V_{EB})$ plot to obtain an intercept I_i. This experimental method determines the currents original name (the intercept current) and subscript. However, it is now normally refered to as I_S and will be refered to as I_S in later chapters of this thesis. From equation 3.7 we can obtain an expression for the emitter terminal current, under the bias conditions shown in figure 3.4.

$$I_E = -\left(I_C + \frac{I_C}{\beta_F}\right) = -I_S\left(1 + \frac{1}{\beta_F}\right)\left(exp\left[\frac{q(-V_{EB})}{kT}\right] - 1\right) \qquad (3.10)$$

we need only equations for two terminal currents as the third can always be obtained from their difference ($-I_E = I_B + I_C$). β_F represents the forward current gain in the common emitter configuration.

If next we operate the transistor in reverse fashion, figure 3.5 illustrates the conditions required. From our previous analysis we can write :

$$I_E = qAD_n\left(\frac{n_B(X_B) - n_B(0)}{X_B}\right) = I_S\left(exp\left[\frac{q(-V_{CB})}{kT}\right] - 1\right) \qquad (3.11)$$

and following equation 3.10. Thus,

$$I_C = -\left(I_E + \frac{I_E}{\beta_R}\right) = -I_S\left(1 + \frac{1}{\beta_R}\right)\left(exp\left[\frac{q(-V_{CB})}{kT}\right] - 1\right). \qquad (3.12)$$

The comparison of equations 3.7 and 3.11 reveal that the BJT is reciprocal in the common base mode. That is, a voltage V_{EB} applied to the left hand port will cause a current I_C at the short circuited right hand port. Conversely, an equal voltage V_{CB} applied at the right hand port will cause an equal current I_E at the short circuited left hand port. This is the condition of repricocity. It would be credible to assume that this condition arises from the use of an idealized 1-D BJT for this analysis. However, in an appendix to their original paper Ebers and Moll showed this result to be virtually independent of device geometry. Later work showed that the reciprocity condition was true for base regions with non-uniform doping profiles [10]. Ebers and Moll had found that reciprocity in the common base BJT is a principle of great generality. They also made another key observation. In equation 3.10 the

emitter current is linearly related to the boundary minority carrier density value $n_B(0)$. Likewise in reverse operation, I_E is linearly related to $n_B(X_B)$. From the principle of superposition developed for linear systems we can obtain a current equation for arbitrary bias values V_{EB} and V_{CB}, Thus the BJT base region is a linear system with its currents varying linearly with its boundary minority carrier density values. Hence, we can obtain a valid two term expression for I_E. However, the boundary density values are related in a grossly non-linear manner to the applied voltages V_{EB}, V_{CB} through the law of the junction. This principle (which requires a linear system) then yields expressions for current which exhibit a highly non-linear dependence on applied voltage. The superposition based expression is

$$I_E = -I_S\left(1 + \frac{1}{\beta_F}\right)\left(exp\left[\frac{q(-V_{EB})}{kT}\right]-1\right) + I_S\left(exp\left[\frac{q(-V_{CB})}{kT}\right]-1\right) \qquad (3.13)$$

and also for the terminal current I_C

$$I_C = I_S\left(exp\left[\frac{q(-V_{EB})}{kT}\right]-1\right) - I_S\left(1 + \frac{1}{\beta_R}\right)\left(exp\left[\frac{q(-V_{CB})}{kT}\right]-1\right). \qquad (3.14)$$

These two simple equations provide the basis for what has been shown to be a very general model. Time has proven this model to be a fundamental tool for the analysis of the BJT. Figure 3.6 shows the diagrammatical implementation of the model.

3.3.2 The EM2 model

The EM2 model is, in simplest terms, an improvement of the original Ebers Moll model described above. The improvement is made by providing first order modelling of charge storage effects, and more importantly for our purposes, a more accurate dc representation of the device. A brief description of the effects on dc characterisation of the model is given below. Figure 3.7 shows the basic additions which make the EM2 model.

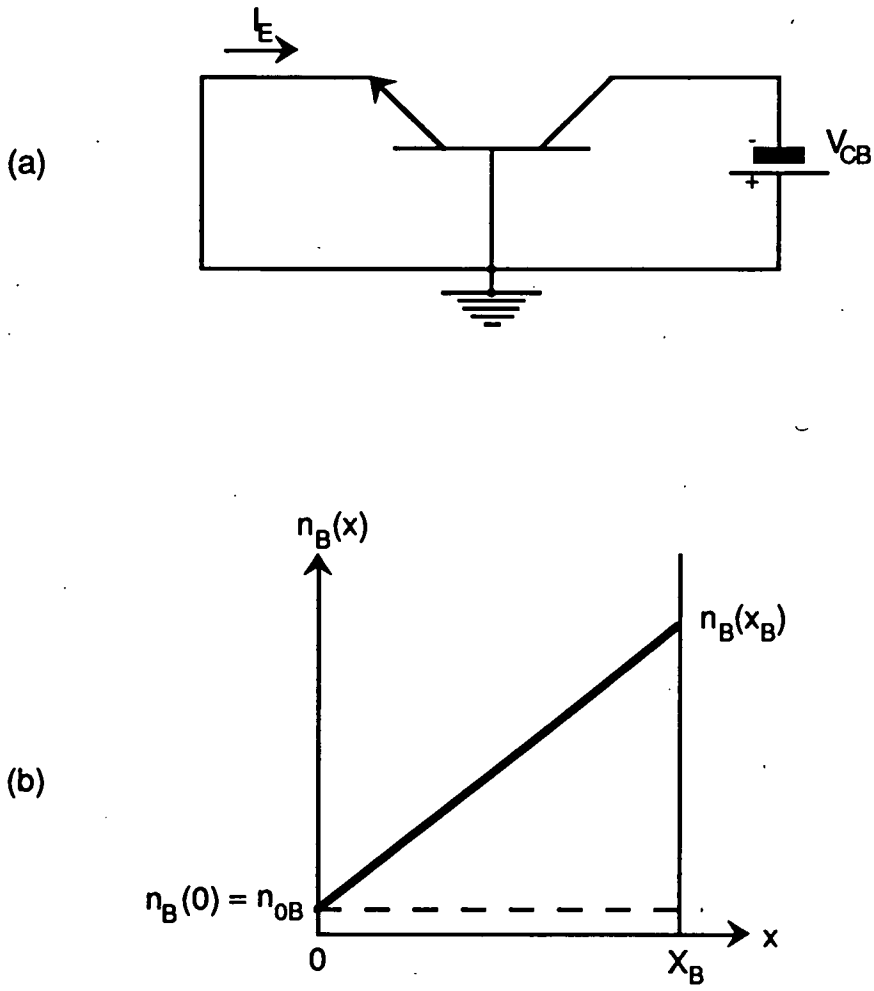


Figure 3.5 Reverse common-base operation. (a) Bias configuration with emitter junction short-circuited and collector junction forward-biased. (b) Base-region minority-electron profile corresponding to conditions in (a).

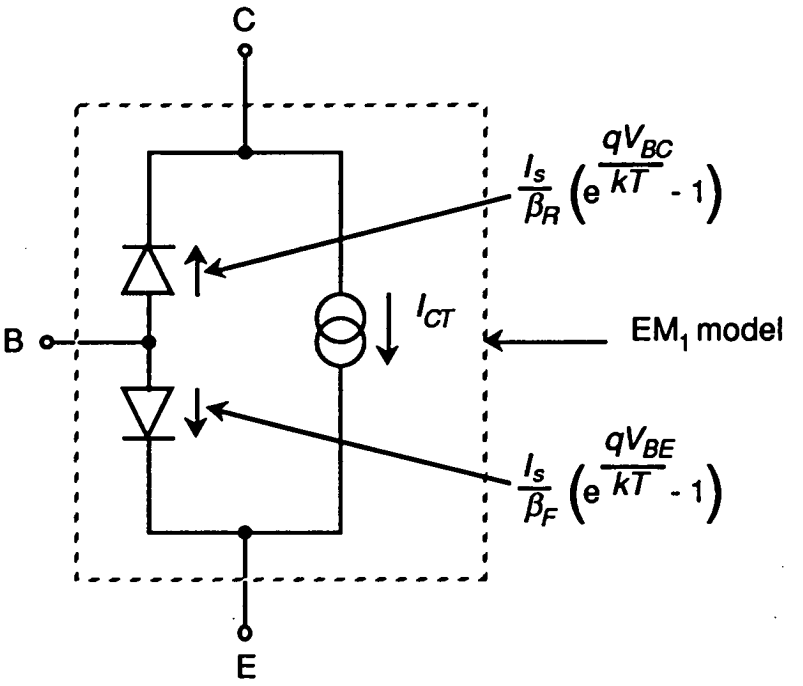


Figure 3.6 Diagramatical implementation of the EM_1 model.

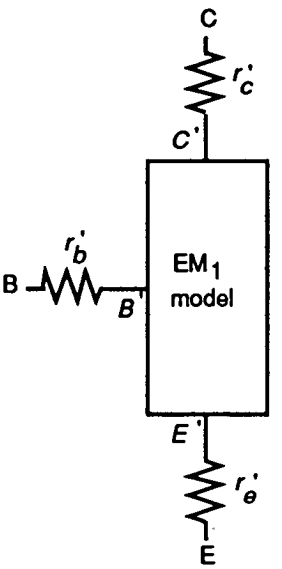


Figure 3.7 Improved DC characterisation of the EM_2 by the inclusion of three constant resistors (r'_e, r'_c and r'_b).

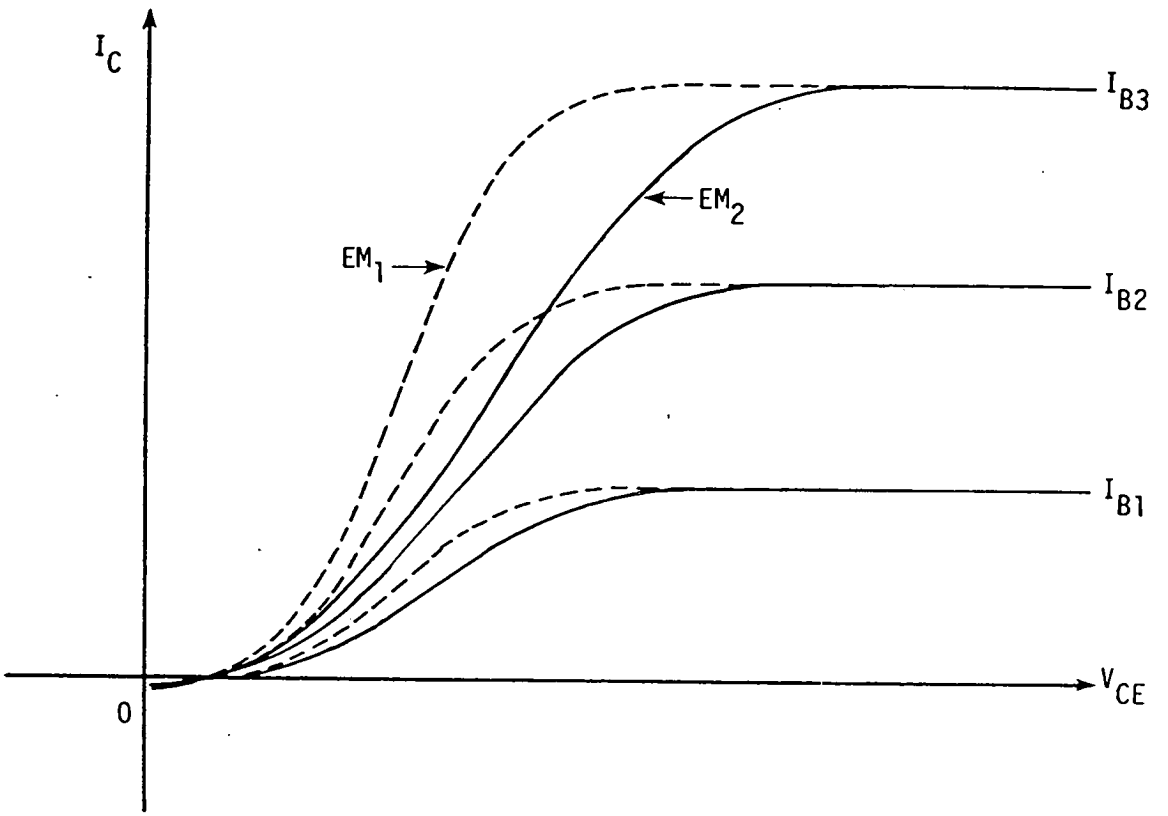


Figure 3.8 The effect of r'_c on the I_C vs. V_{CE} characteristics. Dashed lines represent the EM_1 model ($r'_c = 0$). Solid lines represent the EM_2 model.

The additions of importance for dc analysis are the three fixed resistances r'_c, r'_e and r'_b. The effect of an added collector resistance r'_c in the model can be seen in figure 3.8. Although the EM2 models this parameter as a constant it will be shown later that r'_c varies with collector current and base-collector voltage.

In the modern BJT the emitter is the most heavily doped region of the transistor. This increases gain and improves emitter efficiency [11]. For discrete devices the dominant factor in the overall r'_e is the contact resistance and is of the order of 1Ω. The effect of r'_e is to reduce the applied voltage V_{EB} by a term $I_E r'_e$ at the emitter-base junction. This effect on V_{EB} is equivalent to a base resistance of $(1 + \beta_F)r'_e$, hence r'_e affects I_C and I_B. These effects are illustrated in figure 3.9. The term r'_b is introduced to the model as a constant parameter. However, effects like crowding and the finite value of r'_e make it a difficult parameter to measure accurately [12, 13].

3.3.3 The Gummel Poon Model

To examine the Gummel-Poon model we begin with the Ebers Moll model derived above. Equations 3.13 and 3.14 can be expressed as

$$I_E = I_S\left[\exp\left(\frac{V_{CB}}{V_t}\right) - \exp\left(\frac{V_{EB}}{V_t}\right)\right] + \frac{I_S}{\beta_F}\left[\exp\left(\frac{V_{EB}}{V_t}\right) - 1\right] \tag{3.15}$$

$$I_C = I_S\left[\exp\left(\frac{V_{CB}}{V_t}\right) - \exp\left(\frac{V_{EB}}{V_t}\right)\right] + \frac{I_S}{\beta_R}\left[\exp\left(\frac{V_{CB}}{V_t}\right) - 1\right] \tag{3.16}$$

where

$$V_t = \frac{q}{kT} \tag{3.17}$$

hence

$$I_B = \frac{I_S}{\beta_F}\left[\exp\left(\frac{V_{EB}}{V_t}\right) - 1\right] + \frac{I_S}{\beta_R}\left[\exp\left(\frac{V_{CB}}{V_t}\right) - 1\right]. \tag{3.18}$$

The model proposed by Gummel and Poon [14] modified the relatively straightforward Ebers Moll equations to incorporate three prominent

second order effects.

1. Recombination in the emitter-base space charge layer at low emitter base bias.

2. The current gain decrease experienced under high current conditions.

3. Effects of space charge widening (the Early effect) on the linking current between the emitter and the collector.

Figure 3.10 illustrates the results of these second order effects on the BJT characteristics when biased in the active mode.

To model the effect of recombination in the emitter-base space charge layer Gummel and Poon added four new parameters to the Ebers Moll model. I_{SE}, n_E, I_{SC} and n_C. These four parameters helped define the base current in terms of ideal and non-ideal diode currents (chapter 2).

$$I_B = \frac{I_S}{\beta_F}\left[exp\left(\frac{V_{EB}}{V_t}\right)-1\right] + I_{SE}\left[exp\left(\frac{V_{EB}}{n_E V_t}\right)-1\right]$$

$$+ \frac{I_S}{\beta_R}\left[exp\left(\frac{V_{CB}}{V_t}\right)-1\right] + I_{SC}\left[exp\left(\frac{V_{CB}}{n_C V_t}\right)-1\right] \qquad (3.19)$$

Both the high current effect (2) and the Early effect (3) were incorporated into the model by modifying the value of the linking current I_S. We have seen from equation 3.9 that I_S is fixed by the net impurity in the active base region. If we represent the total base charge as Q_{BT} then

$$I_S = J_S A_E = \frac{q^2 A_E^2 n_i^2 \tilde{D}_n}{Q_{BT}}. \qquad (3.20)$$

In the Gummel-Poon model Q_{BT} is composed of components having bias dependance. There is the "built-in" base charge

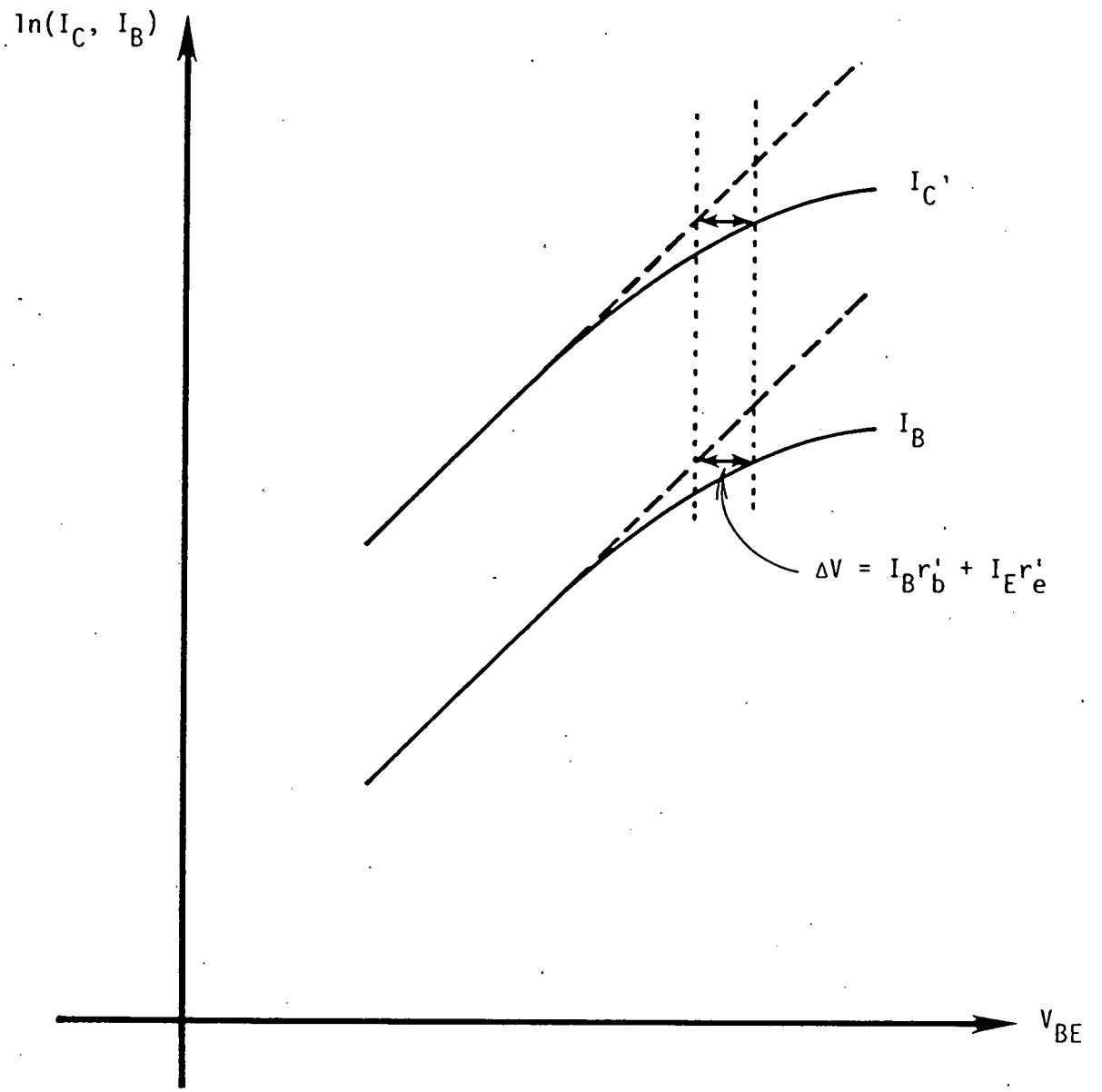


Figure 3.9 The effect of r'_b and r'_e on the $\ln(I_C)$ and $\ln(I_B)$ vs. V_{BE} characteristics of the EM_2 model. (Note that EM_2 model neglects other high-level effects which are treated in higher-order models.)

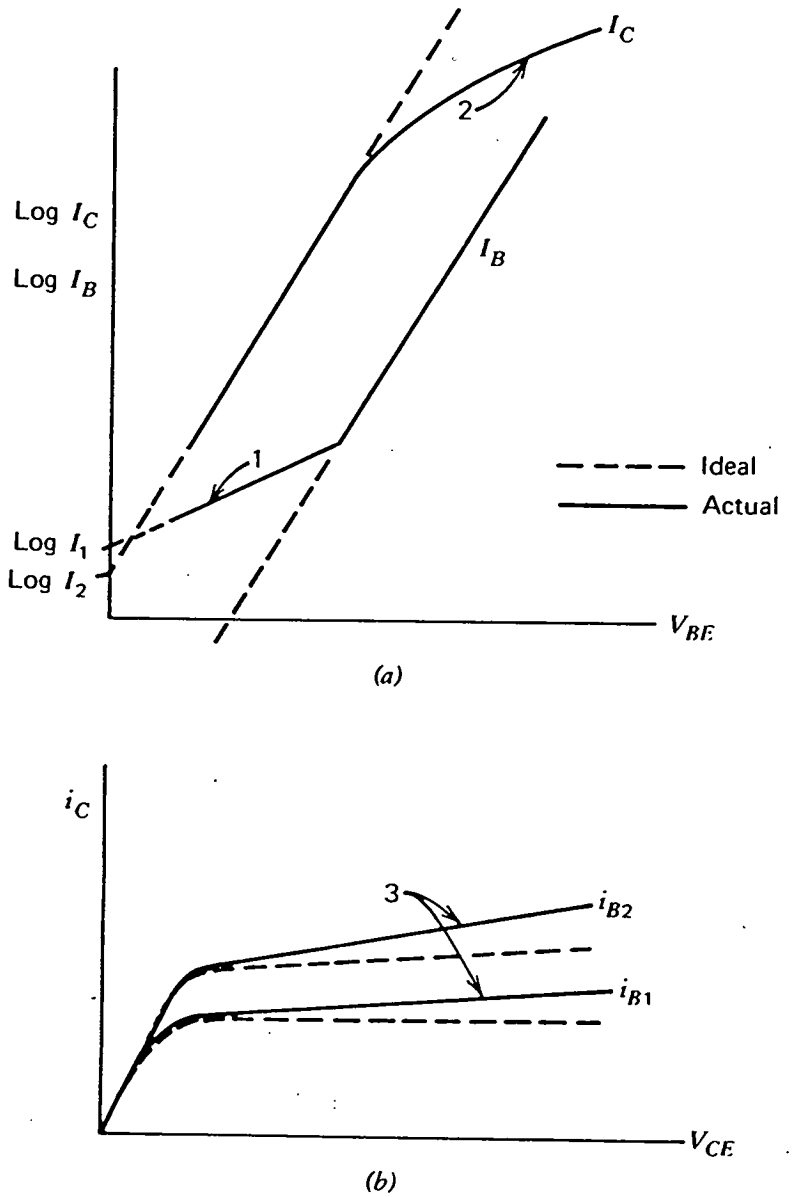


Figure 3.10 The result of second order-effects on bipolar transistor-characteristics in the active mode. The numbers on the figures refer to the effects enumerated in the text. The base current extrapolated to zero base-emitter voltage is I_i.

$$Q_{BO} = qA_E \int_0^{X_B} N_A(x)dx . \tag{3.21}$$

The model includes terms for emitter and collector charge storage (Q_{VE} and Q_{VC}) as well as charge associated with forward and reverse injection of base minority carriers. These are all summed to represent Q_{BT}

$$Q_{BT} = Q_{BO} + C_{je}V_{EB} + C_{jc}V_{CB}\frac{A_E}{A_C} + \frac{Q_{BO}}{Q_{BT}}\tau_F I_S\left[exp\left(\frac{V_{EB}}{V_t}\right)-1\right]$$

$$+ \frac{Q_{BO}}{Q_{BT}}\tau_R I_S\left[exp\left(\frac{V_{CB}}{V_t}\right)-1\right]. \tag{3.22}$$

By defining several parameters equation 3.22 was put into a more manageable format

$$q_b \equiv \frac{Q_{BT}}{Q_{BO}}; \qquad I_{KF} \equiv \frac{Q_{BO}}{t_F}$$

$$I_{KR} \equiv \frac{Q_{BO}}{t_R}; \qquad |V_A| \equiv \frac{Q_{BO}}{C_{jc}}\frac{A_C}{A_E}$$

$$|V_B| \equiv \frac{Q_{BO}}{C_{je}}. \tag{3.23}$$

The key variable, total base charge Q_{BT} is normalised in equation 3.23 to Q_{BO}, and its dimensionless counterpart is designated as q_b. The two charge control time constants τ_F and τ_R define "knee currents" I_{KF} and I_{KR}.

Equation 3.22 can be written in terms of these normalised parameters

$$q_b = q_1 + \frac{q_2}{q_b} \tag{3.24}$$

where q_1 and q_2 are auxiliary variables as defined by

$$q_1 = 1 + \frac{V_{CB}}{|V_A|} + \frac{V_{BE}}{|V_B|}$$

$$q_2 = \frac{I_S}{I_{KF}}\left[exp\left(\frac{V_{EB}}{V_t}\right) - 1\right] + \frac{I_S}{I_{KR}}\left[exp\left(\frac{V_{CB}}{V_t}\right) - 1\right]. \qquad (3.25)$$

The importance of second order effects is highlighted by these two variables. If the Early effect is negligible then q_1 will approach unity. If high level injection effects are not important then q_2 will be small.

Thus base-width modulation effects have been modeled through the introduction of the two Early voltages while high level effects are specified through the knee currents I_{KF} and I_{KR}. In summary, the Gummel-Poon model requires three variables I_S, β_F, and β_R for the basic Ebers-Moll model. It uses four more, I_1, I_2, n_e and n_c to model space-charge-region recombination effects. The Ebers-Moll parameters will be valid in the mid bias range, where high level effects are not present.

The base-width and majority-charge modulation are modelled by specifying a parameter q_b that depends on four additional variables I_{KF}, I_{KR}, V_A and V_B. Thus the overall model is specified by 11 parameters plus the temperature. It is the SPICE/TECAP implementation of this model that will be used for later characterisation of parasitic bipolar devices. There are however several different naming conventions. These are noted in subsequent chapters.

3.4 The MOSFET model

The Metal Oxide Silicon Field Effect Transistor (MOSFET) model incorporated in SPICE/TECAP (Transistor Electrical Characterisation and Analysis Program) actually consists of three different models of varying complexity and accuracy. The model used to characterise MOSFET devices for this experiment was the level 2 model. A brief description of the level 1 and level 2 models is given below.

The level 1 model is based on the Schichman-Hodges model. The level 2 model is a more advanced version of the Schichman-Hodges model which can use either electrical or process type parameters. The second order effects such as channel length modulation are also included in this model. A more detailed account of the level 2 model is given in [15].

3.4.1 The Level 1 Model

The level 1 model allows the user to input device data in the form of electrical parameters which allow direct calculation of the device's performance.

The drain current (I_D) is a function of terminal voltages and the zero bias threshold voltage (V_{TO}). V_{TO} is derived at the onset of strong inversion and marks the point where the device starts conducting if the weak inversion current is ignored. The actual turn-on voltage is related to V_{TO} by

$$V_{ON} = V_{TO} + \gamma\sqrt{\phi - V_{BS}} - \gamma\sqrt{\phi}\ . \tag{3.26}$$

The equation controlling I_D depends on the region of operation of the MOSFET. If $V_{GS} < V_{ON}$, then

$$I_D = 0 \qquad \textit{(cutoff region).} \tag{3.27}$$

If $V_{GS} \geq V_{ON}$ and $V_{DS} < V_{GS} - V_{ON}$, then

$$I_D = K_P(1+\lambda V_{DS})\left(\frac{W}{L-2L_D}\right)V_{DS}\left(v_{GS} - V_{ON} - \frac{V_{DS}}{2}\right) \qquad \textit{(linear region).} \tag{3.28}$$

If $V_{GS} \geq V_{ON}$ and $V_{DS} \geq V_{GS} - V_{ON}$, then

$$I_D = \frac{K_P}{2}(1+\lambda V_{DS})\left(\frac{W}{L-2L_D}\right)(v_{GS} - V_{ON})^2 \qquad \textit{(saturation region).} \tag{3.29}$$

3.4.2 The Level 2 Model

The level 2 model is essentially the level 1 model with many modifications that model second order effects in small geometry devices. Some of the major second order effects that are modeled are back-gate bias and short or narrow channel effects on V_{ON}, saturation due to limited velocity, and finite voltage dependent output conductance. In addition, surface field dependent mobility, weak inversion conduction, variation of all quantities with temperature, and a charge controlled model of regenerative effects are included. The additional equations for

the level 2 model are presented below.

$$C_{ox} = \frac{\varepsilon_{ox}}{T_{ox}} \tag{3.30}$$

$$K_P = \mu_o C_{ox}(10^{-4}) \tag{3.31}$$

$$\phi = \frac{2kT}{q} ln\left(\frac{N_{sub}}{n_i}\right) \tag{3.32}$$

$$\gamma = \frac{\sqrt{2\varepsilon_{si} q N_{sub}}}{C_{ox}} \tag{3.33}$$

3.4.2.1 Effective Channel Length

The effective channel length compensates for diffusion effects in the length of the channel. L_{eff} is given as

$$L_{eff} = L - 2L_D. \tag{3.34}$$

L_D is the lateral diffusion coefficient, this is referred to as ΔL in later chapters (6,7). In all subsequent equations L will be L_{eff}. For a transistor of width W and length L

$$\beta = K_P \frac{W}{L} \tag{3.35}$$

$$C_{OX} = C_{ox} WL \tag{3.36}$$

3.4.2.2 Threshold Voltage

According to the size of each transistor, SPICE modifies V_{TH} due to the following effects.

- An increase of the bulk-to-source voltages increases the depletion charge which increases V_{TH}.

- In a short channel device some of the depletion charge in the bulk terminates the electric field of the drain and source

junctions. This helps to lower the gate-to-bulk electric field thus lowering V_{TH}.

- The amount of charge underneath the gate is depleted by the drain junction field which lowers V_{TH}.

These effects are modeled using the following equations

$$V_{bin} = V_{TO} - \gamma\sqrt{\phi} + F_N(\phi - V_{BS}) \tag{3.37}$$

$$V_{ON} = V_{bin} + \gamma_d\sqrt{\phi - V_{BS}} + n\frac{kT}{q} \tag{3.38}$$

where

$$n=1 + q\frac{N_{FS}}{C_{OX}} - \frac{\delta}{\delta V_b}\left[\gamma_d\sqrt{\phi - V_{BS}}\right] + F_N \tag{3.39}$$

and when $x_j \neq 0$

$$\gamma_d = \gamma\left\{1 - \frac{x_j}{2L}\left[\sqrt{1 + \frac{2W_s}{x_j}} - 1\right] - \frac{x_j}{2L}\left[\sqrt{1 + \frac{2W_d}{x_j}} - 1\right]\right\} \tag{3.40}$$

where

$$W_d = x_d\sqrt{\phi + V_{DS} - V_{BS}} \text{ and } W_s = x_d\sqrt{\phi - V_{BS}}\ . \tag{3.41}$$

3.4.2.3 Effective mobility

The level 1 model treats mobility (μ) as a constant which leads to I_D vs V_{GS} curves that are straight lines. The level 2 model more accurately models the reduction in channel mobility as a function of the applied electric field from the gate.

$$\mu_{eff} = \mu_o\left[\frac{U_{crit}\varepsilon_{si}}{C_{ox}(V_{GS} - V_{ON} - U_{tra}V_{DS})}\right]^{U_{exp}} \tag{3.42}$$

The parameter U_{tra} is a TECAP addition to the model to provide more flexibility. However, in subsequent characterisations the parameter will be set to zero. This makes the model equivalent to that in SPICE.

3.4.2.4 Saturation Voltage

For $V_D \leq V_G - V_T$ and $V_G \geq V_T$, the MOSFET is in the ohmic or linear region of operation. A plot of I_D vs V_D in this region shows characteristic curves that can be modeled as parabolas. However, as V_D approaches $V_G - V_T$, the depletion region near the drain *"pinches off"* the channel and the drain current approaches a constant value. As described in the next section, channel length modulation causes the slope of V_D vs I_D to be > 0. SPICE/TECAP can model this effect in two ways. The first technique uses the classical Grove-Frohman equation which models V_{Dsat} as a function of drain voltage and pinch off. Thus,

$$V_{Dsat} = \frac{V_{GS} - V_{bin}}{\eta} + \frac{1}{2}\left(\frac{\gamma_d}{\eta}\right)^2 \left\{ 1 - \sqrt{1 + 4\left(\frac{\eta}{\gamma_d}\right)^2 \left[\frac{V_{GS} - V_{bin}}{\eta} + \phi - V_{BS}\right]} \right\}. \quad (3.42)$$

In short channel MOSFETs, the electrons in the channel reach a maximum velocity before the depletion region around the drain limits current. This maximum velocity causes V_{Dsat} to be reached faster than the Grove-Frohman model. SPICE/TECAP can model this effect using Baum's theory of scattering velocity saturation. Unfortunately, this model requires the simultaneous solution of two nonlinear equations which greatly increase the computation time, The solution that SPICE/TECAP uses to solve this computational problem is rather involved [15]. Thus, the velocity saturating model is defined such that

if $V_{MAX} > 0$, then $V_{Dsat} = V_{DS}$ when $V_{MAX} = \frac{I_{Dsat}}{WQ_{chan}}$.

3.4.2.5 Channel Length Modulation

MOS transistors exhibit a finite drain-to-source conductance in the saturation region somewhat analogous to the Early effect in the bipolar transistor. This conductance is primarily due to the effect of V_{DS} modifying the depletion region around the drain which in turn varies

the effective length. This leads to the following equations for the channel length modulation parameter.

For $V_{MAX} \leq 0$

$$\lambda = \frac{x_d}{LV_{DS}} \sqrt{\frac{V_d - V_{Dsat}}{4} + \sqrt{1 + \left(\frac{V_d - V_{Dsat}}{4}\right)^2}} \quad . \tag{3.43}$$

For $V_{MAX} > 0$

$$\lambda = \frac{x_d}{LV_{DS}\sqrt{N_{eff}}} \left\{ \sqrt{\left(\frac{x_d V_{MAX}}{2\mu_{eff}\sqrt{N_{eff}}}\right)^2 + V_{DS} - V_{Dsat}} - \frac{x_d V_{MAX}}{2\mu_{eff}\sqrt{N_{eff}}} \right\} . \tag{3.44}$$

3.4.2.6 Channel Shortening at Punch-Through

A possible problem with a short channel device is that at a high V_{DS} punch-through occurs. In order to prevent the channel length from going negative in the last two equations for λ, the effective channel length L_{eff} must be kept larger than the zero-bias depletion layer width W_B; where $W_B = x_d\sqrt{P_B}$.

$$L_{eff} = L(1 - \lambda V_{DS}) \qquad for\ L_{eff} \geq W_B \tag{3.45}$$

$$L_{eff} = \frac{W_B}{1 + \frac{\lambda V_{DS} L - L + W_B}{W_B}} \qquad for\ L_{eff} < W_B \tag{3.46}$$

3.4.2.7 Mobility and Channel Modulation

As a second order effect, the applied electric field from the gate modifies the carrier mobility in the channel. As noted above, the effective channel length is modified by V_{DS}. These effects are incorporated with a parameter defined as β.

$$\beta_1 = \beta \frac{\mu_{eff} L}{\mu_o L_{eff}} \tag{3.47}$$

Another parameter is *Body*, where

$$Body = (\phi + V_{DS} - V_{BS})^{3/2} - (\phi - V_{BS})^{3/2}. \tag{3.48}$$

3.4.2.8 Drain Current Equations

The drain current equations for the level 2 model include the above improvements in modelling accuracy. They are given here:

if $V_{GS} < V_{ON}$, then

$$I_D = \beta_1 \left\{ \left(V_{ON} - V_{bin} - \eta \frac{V_{DS}}{2} \right) V_{DS} - \frac{2}{3}\gamma_d Body \right\} exp\left(\frac{V_{GS} - V_{ON}}{\frac{nkT}{q}} \right) \quad (subthreshold) \tag{3.49}$$

if $V_{GS} > V_{ON}$ and $V_{DS} \leq V_{Dsat}$, then

$$I_D = \beta_1 \left\{ \left(V_{gs} - V_{bin} - \eta \frac{V_{DS}}{2} \right) V_{DS} - \frac{2}{3}\gamma_d Body \right\} \quad (linear) \tag{3.50}$$

if $V_{gs} > V_{ON}$ and $V_{DS} > V_{Dsat}$, then

$$Body_{sat} = (\phi + V_{Dsat} - V_{bs})^{3/2} - (\phi - V_{bs})^{3/2}$$

$$I_D = \beta_1 \left\{ \left(V_{gs} - V_{bin} - \eta \frac{V_{Dsat}}{2} \right) V_{Dsat} - \frac{2}{3}\gamma_d Body_{sat} \right\} \quad (saturation). \tag{3.51}$$

The above equations represent the model that will be used in later chapters to characterise the dc performance of the MOS devices fabricated for this project.

3.5 TECAP Operation and Hardware Requirements

The HP 94445A TECAP software links the system controller to HP semiconductor measurement equipment. TECAP (Transistor Electrical Characterisation and Analysis Program) transforms measured test

data into transistor model parameters. These model parameters are used as device measures for device evaluation and comparison. TECAP combines measurement, parameter extraction and device model simulation all in one program. A typical application of TECAP is illustrated in figure 3.11.

In most cases TECAP obtains the required device model parameters by performing the following functions (figure 3.12).

1. Measures the device performance characteristics under specified test conditions.

2. Extracts the specified model parameters from the measurement data for the active model.

3. Simulates the performance of the specified model using all of the parameters in the active model parameter table.

4. Compares the measured and simulated data to determine if the model parameters now stored in the model parameter table meet the users accuracy requirements. If they do not, additional extraction and simulation cycles can be performed to optimize the parameters.

5. If the model parameters are acceptable, they are formatted and stored in a file that can be transferred to the spice system.

3.5.1 Hardware Requirements

An HP 9836C desktop computer with at least 2 megabytes of memory is used as the system controller. Basic mass storage is provided by the HP 9133D winchester/3.5 inch microfloppy system. The HP-IB provides communication between the system controller and all instruments in the system. All instruments are connected in parallel to the HP-IB without any custom interfacing. Figure 3.13 shows the hardware configuration used in the EMF system. The HP4145A semiconductor parameter analyser is a fully automatic, high performance instrument for measuring the DC current and voltages of the IC devices. This instrument is equipped with four programmable

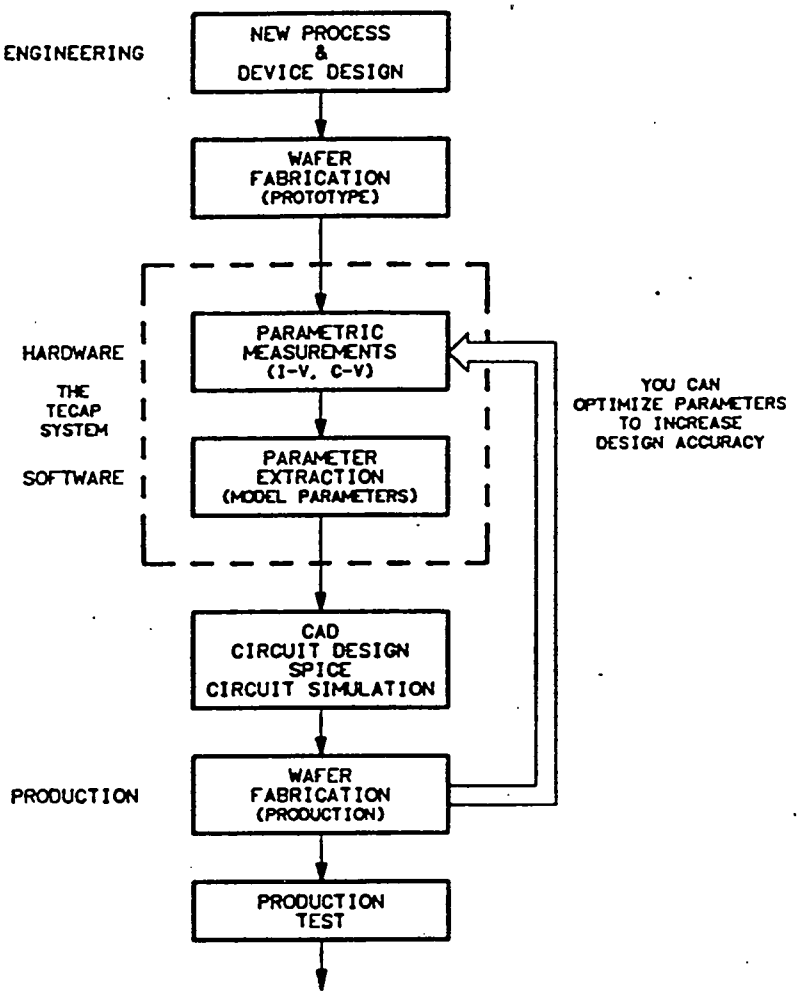


Figure 3.11 Typical TECAP system application.

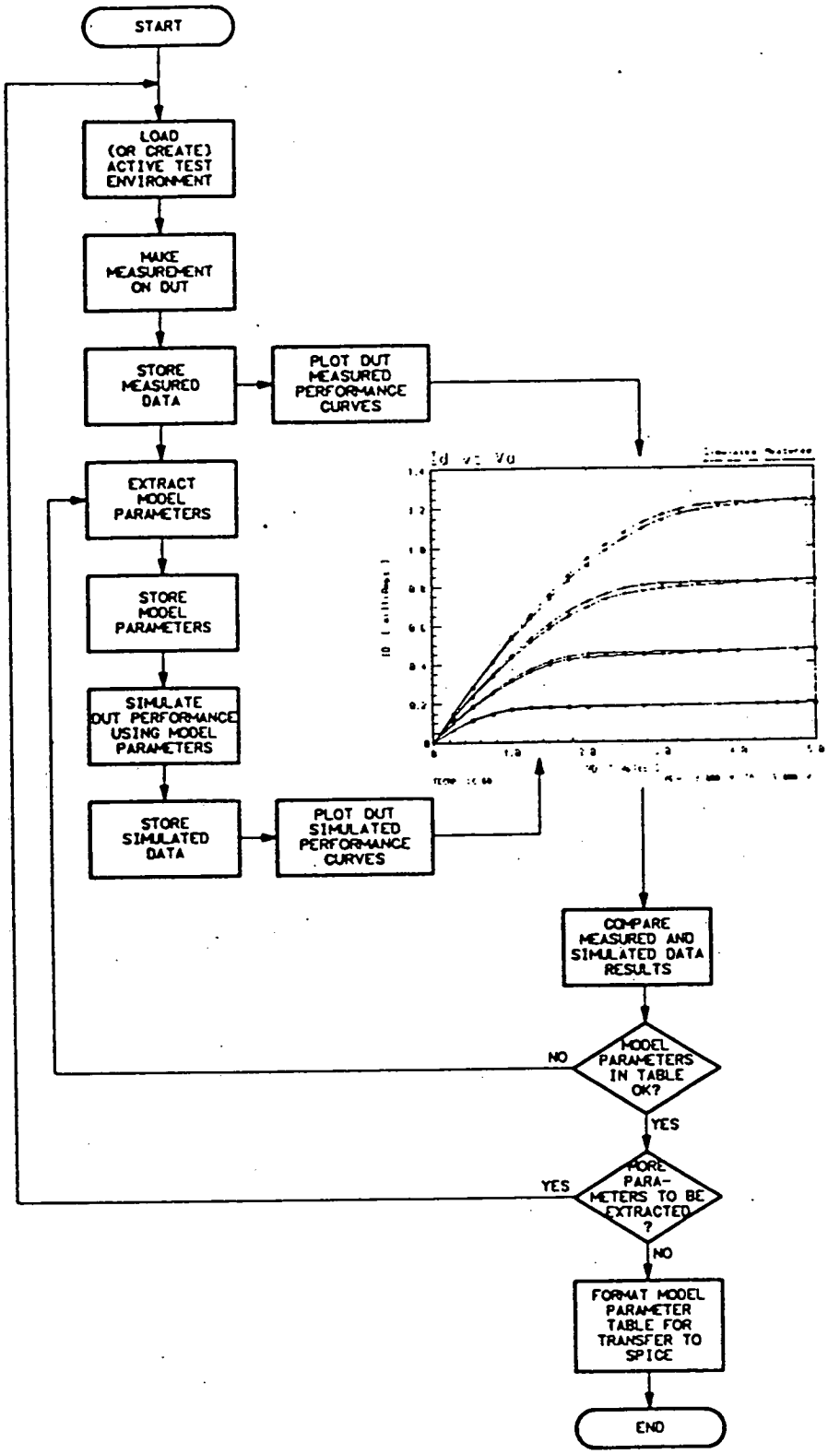


Figure 3.12 System operational flowchart.

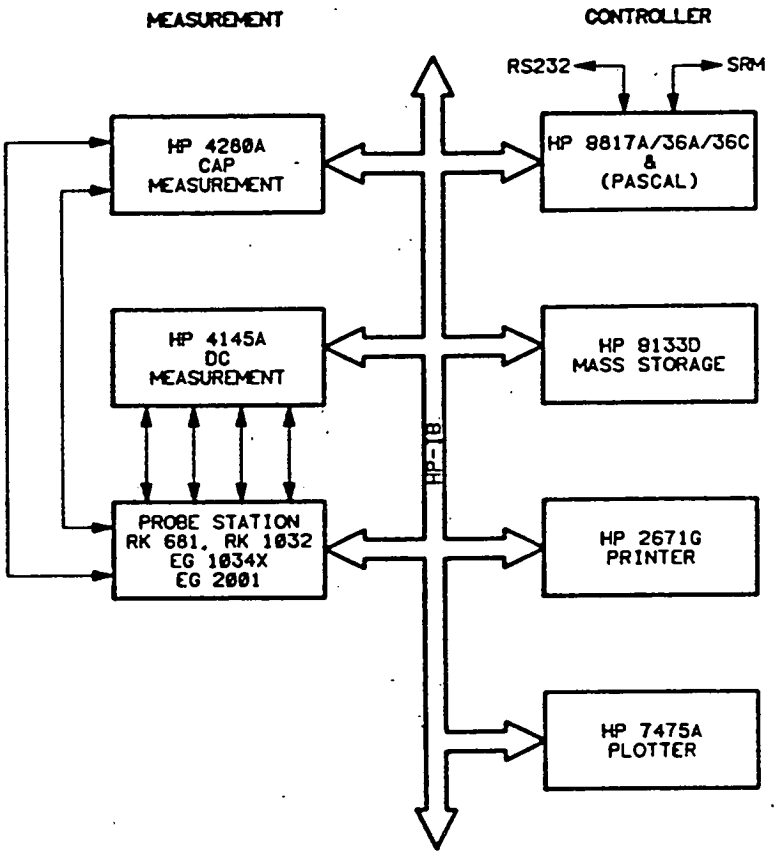


Figure 3.13 Hardware configuration used for parameter extraction.

Stimulus/Measurement Units (SMU). Each SMU can be programmed to function as a voltage source and current monitor or as a current source and voltage monitor. In this mode of operation, each SMU has a range of 1 picoamp to 100 milliamps and 1 mV to 100 Volts.

3.5.2 TECAP Models

TECAP utilises the UCB-BIPOLAR model for bipolar transistor characterisation. This is an exact copy of the UC Berkley 2G.6 version of the SPICE program. This model is equivalent to that described in section 3.3.3. For MOS characterisation TECAP uses the UCB-MOSFET model for MOS device characterisation. This is an exact copy of the model used in the UC Berkley 2G.5 and 2G.6 models of the SPICE program except for two parameters.

The parameter UTRA, which has been removed from the SPICE model is revitalized in the TECAP model. This parameter models the effect of drain voltage on mobility reduction.In order to obtain the same results as SPICE this parameter is set to zero. The parameter WD is added to the TECAP model to provide the scaling with width (especially for the narrow devices). The effective channel width is then calculated as $W_{eff} = W - 2xWD$. The parameter is extracted in TECAP but cannot be used directly with SPICE. The WD effect, however can be accounted for in SPICE by subtracting 2xWD from the channel width of the narrow channel transistors.

The UCB-MOSFET model is made up of three separate models; Level 1, Level 2 and Level 3. All three levels are implemented in TECAP but only the level 2 parameters are extracted by the predefined TECAP commands. As a result it is the level 2 model which is used in subsequent MOS device characterisations.

References

[1] George L. Schnable,"Failure Mechanisms in LSI Circuits," *IEEE Trans.*, Ed 16, No. 4

[2] Martin Buehler,"Micro Electronic Test Chips for VLSI Electronics," *Academic press*, 1983.

[3] Bob Johnston,"Process Control in Wafer Fabrication,"*Solid State Technology*, May 1986, pp195-199.

[4] Glenn Evans,"Semiconductor Parametric Testing: Yesterday, Today, and Tomorrow,"*Microelectronic Manufacturing and Testing*, Jan 1983.

[5] Glenn Evans,"Parametric Testing for LSI/VLSI,"*Microelectronic Manufacturing and Testing*, Jan 1983.

[6] Chris Chrones,"Parametric testers evaluate wafer processing,"*EDN*, April 1981.

[7] David Angst and Joe Domitrowich,"Trends in Parametric Test Systems," *Semiconductor International*, Sep 1987.

[8] J. J. Ebers and J. L. Moll,"Large Signal Behaviour of Junction Transistors," *Proc. IRE 42*, p1761, 1954.

[9] H. K. Gummel,"Measurement of the Number of Impurities in the Base Layer of a Transistor," *Proc. IRE*, pp 834, 1961.

[10] B. L. Hart,"Direct Verification of the Ebers-Moll Repricocity Condition," *Int. J. Electron.* 31, p. 293, 1971.

[11] A.B. Phillips,"Transistor Engineering," *McGraw-Hill*, Section 7.5, 1962.

[12] W. M. C. Sansen and R. G. Meyer,"Characterisation and Measurement of the Base and Emitter Resistances of Bipolar Transistors," *IEEE J. Solid-State Circuits*, vol. SC-7, pp 492-498, 1972.

[13] J. R. Hauser,"The Effects of Distributed Base Potential on Emitter-Current Injection density and Effective Base resistance for Stripe Transistor Geometries," *IEEE Trans. Electron Devices*, vol ED-11, pp. 238-242, 1964.

[14] H. K. Gummel and H. C. Poon, *Bell syst. Tech. J.*, 49, p 827, 1970.

[15] Andrei Vladimerescu and S. Liu,"The Simulation of MOS Circuits Using SPICE2,"*Electronics Research Laboratory*, No. UCB/ERL M80/7, 1980.

Chapter 4

Test Chip Design, Fabrication And Characterisation

Introduction

To investigate the feasibility of using parasitic structures to monitor a CMOS process, test structures were designed and fabricated. This chapter details the design and processing aspects of two test chips which were created for this purpose.

4.1 Aims of The Test Chip

Test chips must provide useful electrical characterisation information or in the case of vernier-type structures optical information. The measurements taken from a test chip should be repeatable and relatively straight-forward. In the case of process introduction more latitude is available to the test chip designer as the primary aims are somewhat different from that of process control. In this project the test chips were designed as process control tools. The chips were designed in order to provide as much end-of-line process information as possible. The individual structures had to be designed within the constraints of the measuring equipment available to measure their characteristics at the end of the line. More detail is provided at the end of this chapter. In summary, the only testing available was dc electrical characterisation. As discussed in chapter 3 this is typical of most CMOS semiconductor

fabrication lines.

Probably the most important aspect of the test chip was that it had to be fabricated using no extra mask steps or implants. The processing had to conform to the standard CMOS process available. Any structures found to be useful could then be dropped straight into existing test chips. Thus, anyone interested in the technique can incorporate the method in their standard CMOS test chips.

4.2 The EMF 5μm Process

The EMF 5μm process is an n-well CMOS process. The major steps and details of the process are listed here.

The starting material is 14-20 ohm.cm p-type silicon with (100) orientation. The wafers are first cleaned in a sulphuric acid-hydrogen peroxide mix before an initial oxide is grown. The wafers are then masked for n-well implant. The implant is P31+ at a dose of $2x10^{12}$ atoms/cm^2 with an energy of 100keV. The n-well drive-in is 18hrs standard*. For LOCOS isolation, a pad oxide is grown and a silicon nitride layer is deposited. The active areas are patterned and the nitride etched. The wafers are then patterned for a substrate field implant of B11+, dose $7x10^{13}$ atoms/cm^2. Field oxide is grown and the remaining nitride etched off. After a 10% HF dip the gate oxide is grown in a 17 min wet oxidation with HCl at 950°C. A polysilicon layer is then deposited. The polysilicon is phosphorus doped from a solid source. The polysilicon is then deglazed and oxidised in a wet 15 min oxidation. After the gates are patterned the polysilicon is then plasma etched. The n+ source-drain regions are then patterned and implanted. The implant is P31+, dose $7x10^{15}$ atoms/cm^2 with an energy of 90keV. The implant is then annealed for 30 minutes at 950°C. The p+ source-drain regions are then patterned and implanted. The implant is B11+, dose $2x10^{15}$ atoms/cm^2 with an energy of 30keV. The wafers are then given a 15 minute wet oxidation before a pyrolitic oxide is deposited. The pyrolitic oxide is reflowed for 10 minutes. The wafers are oxidised for 15 minutes in a wet oxidation at 950°C. The contacts are patterned and cut. The wafers receive a second reflow before the metal layer is deposited, patterned, etched and sintered.

* Temperature of drive in is 1150°C

4.2.1 Design Rules

The following tables give details of the design rules used in the design of test chips for this experiment.

Rule	Description	Dimensions (μm)
W1	Minimum width of n-well	7.0
W2	Minimum overlap of diffusion island inside well by the n-well	3.5
W3	Minimum separation of n-wells at different potentials.	16.0
W4	Minimum separation of n-wells at the same potential	12.0
W5.	Minimum separation of n-well and n+ diffusion islands outside the well	8.0
W6	Minimum separation between n-well and p+ diffusion island outside the well	11.0

Table 4.1 Layer No 1-n-wells

Rule	Description	Dimensions (μm)
A1	Minimum width of diffusion island	4.0
A2	Minimum separation of diffusions of the same doping type in the same substrate	3.5
A3	Minimum separation of oppositely doped diffusions in the same substrate	6.5

Table 4.2 Layer No 2 Active Areas (Diffusion Islands)

Rule	Description	Dimensions (μm)
PY1	Minimum poly width	3.0
PY2	Minimum poly to poly separation	3.0
PY3	Minimum extension of gate forming poly beyond diffusion island edge	2.5
PY4	Minimum separation between poly gate and diffusion island (transistor source/drain) outer edge	4.0
PY5	Minimum separation between poly interconnect and diffusion island edge	1.5

Table 4.3 Layer No 3 Polysilicon

Rule	Description	Dimensions (μm)
N1	Minimum overlap of n+ implant over diffusion island to be doped n+	2.5
N2	Maximum overlap of n+ implant over diffusion island	12.0
N3	Minimum separation between n+ implant and diffusion islands to be doped p+.(excluding rule N5)	2.5
N4	Minimum separation between separate n+ implant shapes.(otherwise merge shapes)	4.0 or merge = 0
N5	Overlap of n+ implant shape and p+ implant shape when connected to same voltage by metal short	0
N6	Minimum dimension of diffusion island receiving only n+ implant	6.0
N7	Minimum separation between n+ implant and poly gate edge of p-type transistor	6.0

Table 4.4 Layer No 4 N+ Diffusion Implant

Rule	Description	Dimensions (μm)
P1	Minimum overlap of p+ implant over diffusion island to be doped p+	2.5
P2	Maximum overlap of p+ implant over diffusion island	12.0
P3	Minimum separation between p+ implant and diffusion islands to be doped n+.(excluding rule P5)	2.5
P4	Minimum separation between separate p+ implant shapes.(otherwise merge shapes)	4.0 or merge = 0
P5	Overlap of p+ implant shape and p+ implant shape when connected to same voltage by metal short	0
P6	Minimum dimension of diffusion island receiving only p+ implant	6.0
P7	Minimum separation between p+ implant and poly gate edge of p-type transistor	6.0

Table 4.5 Layer No 5 P+ Diffusion Implant

Rule	Description	Dimensions (μm)
C1	Prefered contact size	3 x 3
C2	Maximum length of contact (prefer string of C1 contacts)	9.0
C3	Minimum separation between contacts into same silicon shape	3.0
C4	Minimum poly (interconnect) overlap of contact	2.0
C5	Minimum diffusion island overlap of contact	2.0
C6	Minimum separation between contact.to poly interconnect and diffusion	3.0
C7	Minimum separation between contact to edge diffusion island and poly edge	3.0
C8	Extension of contact bridging p+/n+ into pure doping	4.0
C9	Minimum separation between contact edge and doping interface when using separate contacts	3.0

Table 4.6 Layer 6 Contacts

Rule	Description	Dimensions (μm)
M1	Minimum width of metal	4.0
M2	Minimum separation between metal tracks	4.0
M3	Minimum metal overlap of contacts	2.0
M4	Minimum spacing of metal edges and parallel poly edges for runs in excess of 10 μm	1.5

Table 4.7 Layer 7 Metal

4.3 Initial Design Thoughts

The starting point for the test chip design was the examination of traditional bipolar and JFET designs. From these, structures which can be fabricated in a standard CMOS process can be designed. Because newer bipolar designs incorporate many process improvements, ie epitaxial layers, schottky clamping, pipe emitters and very thin base-widths, the starting point, in the case of the bipolar designs was to examine some typical IC bipolar transistors and extract the salient features. Figures 4.1 a-c show three typical transistors. Figure 4.1a shows a junction-isolated, digital IC transistor with 5μm minimum dimensions. Figure 4.1b shows an amplifying (35 V breakdown) transistor taken from an analogue circuit, once again junction isolated. Figure 4.1c shows an oxide-isolated npn transistor made using a p-type epitaxial layer and polysilicon connections to the base and emitter. Table 4.8 gives the typical design parameters for these transistors [1].

4.3.1 Traditional bipolar and JFET designs

To illustrate the difference in performance of parasitic and traditional devices it is helpful to discuss typical design parameters for

Epitaxial film	Digital Switching Transistor (fig 4.1c)	Junction isolated analogue Device (fig 4.1b)	Oxide isolated analogue device
Thickness	10μm	3.0μm	1.2μm
Resistivity	1Ω.cm	0.3-0.8Ω.cm	0.3-.8Ω.cm
Buried layer			
Sheet resistance	~20 Ω/sq	~30 Ω/sq	
Up diffusion	2.5 μm	1.4μm	3μm
Emitter			
Diffusion depth in base	2.5μm	8μm	25μm
Sheet resistance	5 Ω/sq	12 Ω/sq	30 Ω/sq
Base			
Diffusion depth	3.25μm	1.3μm	.5μm
sheet resistance	100 Ω/sq	200 Ω/sq	600 Ω/sq
Substrate			
Resistivity	10 Ω.cm	10 Ω.cm	~5Ω.cm
Orientation	(111)	(111)	(111)

Table 4.8 Typical Design parameters for three types of BJT.

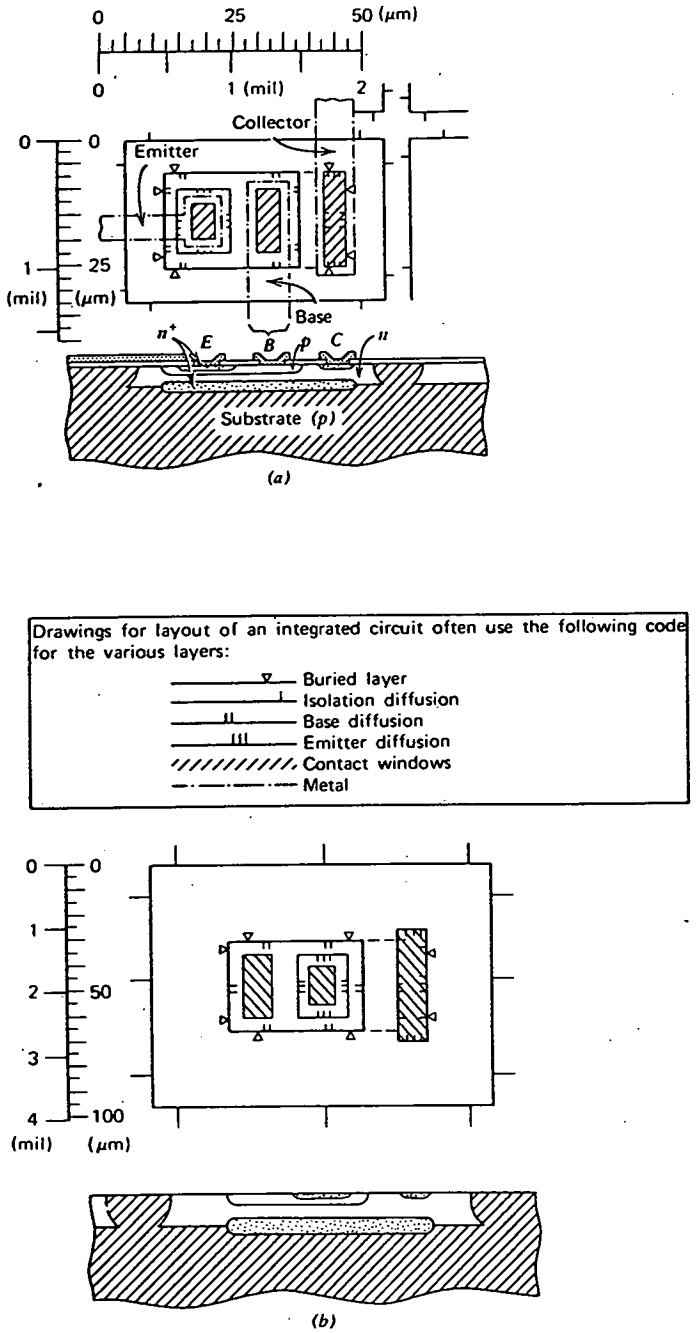


Figure 4.1 Top views (showing dimensions) and cross sections of *(a)* a junction isolated, digital *IC* transistor with 5μm minimum dimensions, *(b)* an amplifying (35V breakdown) transistor. Taken from reference [2].

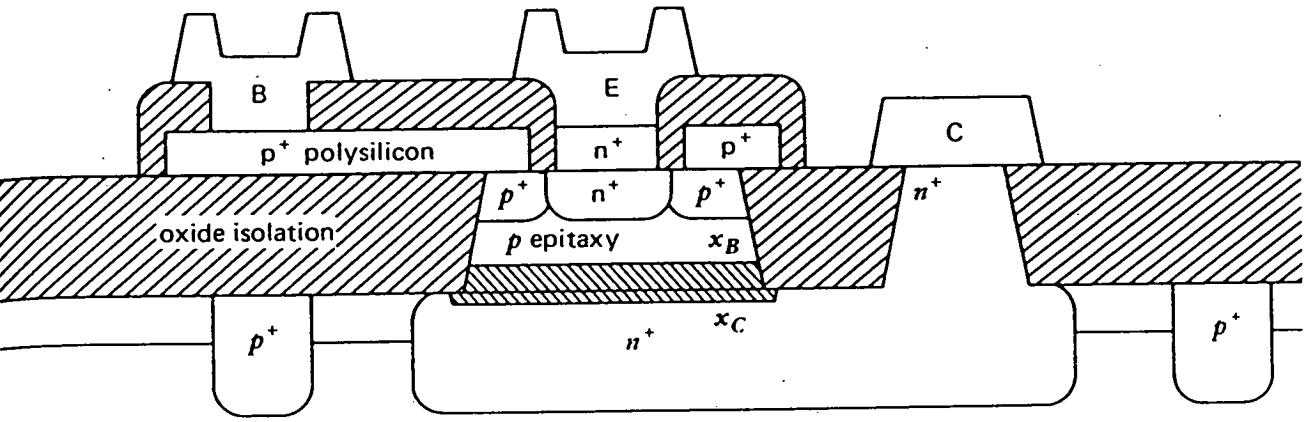


Figure 4.1c. Cross section of an oxide isolated npn transistor made using a p-type epitaxial layer and polysilicon connections to the base and emitter. The collector-base space charge layer extending from x_B to x_C is shown by the (left slanting) cross-hatching. Taken from reference [3].

some traditional devices and the effect these parameters have on device behavior. The transistors shown in figure 4.1 can be split into two broad categories defined by their ultimate use: either amplification or switching. Transistors of either type are nearly always fabricated in an epitaxial layer of relatively high resistivity silicon. Typical values are given in table 4.8. The purpose of this epitaxial layer is to obtain a controlled, lightly doped region of collector material adjacent to the base junction. A buried layer of heavily doped silicon is added to maintain a high punch through voltage.

The lightly doped region allows the collector-base junction to sustain relatively high voltages without breaking down, while the higher doped buried layer region reduces series resistance between the junction and the metallic collector contact. Inclusion of the buried layer reduces series ohmic resistance from the kiloohm range to a few hundred ohms in typical devices. For some applications even this resistance is too much, and an extra processing step is added in which a heavily doped n+ region under the collector contact is diffused until it reaches the buried layer. This extended diffusion region is called a collector plug, and its inclusion reduces series resistance to the order of 10 ohms. The base doping level is ideally very low in order to maximize emitter injection efficiency, gamma, but it cannot be too low (not less than about 5×10^{16} cm^{-3}) without the possibility of a poor metal-semiconductor contact or even surface inversion at the metallic contact to the base. If the doping is too low, series resistance in the base also becomes problematic. The emitter is usually doped very heavily to increase emitter efficiency. When doped above 10^{20} cm^{-3}, however, efficiency again falls because of decreased hole lifetime in the emitter (which increases hole injection from the base) and because of effects associated with the narrowing of the band gap in degenerate silicon [1].

Reducing the base width and decreasing base conductivity are two means of increasing gain, as noted in chapter 2. As the base width is reduced to submicron dimensions it becomes increasingly likely that the collector-base space charge region may reach through to the emitter-base space charge region, completely depleting the base region. This condition, called punch-through results in a highly conductive path from emitter to collector and can lead to damaging currents in a transistor.

JFET fabrication employs two main techniques,

1. Epitaxial layers. These can be used to provide a uniformly doped channel region then implantation is used to provide the gate electrode. This results in a fairly uniform channel depth.

2. Implantation. This is used to make the channel and the gate electrode.

The choice of fabrication techniques is usually determined by whether epitaxy is required for any other devices on the IC. JFETs are usually incorporated in a circuit to obtain a high input resistance, for example, in operational amplifiers.

4.3.2 Feasibility with EMF 5μm process

The primary difference between the EMF 5μm CMOS process and those detailed above is the lack of an epitaxial layer. The CMOS design rules given above are implemented to eliminate or at least reduce the chance of device latch-up. Thus no significant advantage can be offered to the MOS structures to warrant the inclusion of epitaxy in the process. For the bipolar device this means that the collector region (in the case of the vertical bipolar device) is primarily controlled by the silicon starting material, 14-20 Ω.cm p-type silicon. This compares with a 1.0 Ω.cm or less, buried layer resistivity. Thus any bipolar device fabricated with the EMF process will have, by comparison, a large collector resistance.

The EMF process has 3 implants which can be used to form devices. These are the n-well implant and the p+ and n+ source-drain implants. To fabricate bipolar and JFET devices, the n-well implant can be used as the base implant for the bipolar devices, and as the channel implant for JFETs. We have seen above that base implants should be lightly doped but not too lightly ($>5x10^{16}$ cm^{-3}). The n-well implant of $2x10^{12}$ atoms/cm^2 results in a base doping of $1x10^{16}$ cm^{-3}, this doping concentration was calculated using SUPREM3. Simulation of the process is discussed later in the chapter. The resultant dose, from above, is a relatively low dose for a base. The effects are increased base

resistance and reduced gain of any bipolar transistor fabricated with this process. Altough low for the base of a bipolar transistor, this is a relatively high dose for the channel implant of a JFET device. It reduces the channel resistivity which is a major disadvantage for operational JFET devices.

Another concern arising from the use of the n-well implant as a base or channel implant is that of junction depth. With a drive-in time of 18hrs, the n-well junction depth is around 6.5μm. This compares with a typical bipolar base depth of 3.5μm or less. As bipolar gain is proportional to base width, this effect alone drastically reduces the gain of any bipolar device fabricated with the EMF process. As will be shown later the channel depth (n-well junction depth) of a JFET device is directly proportional to the threshold of a JFET device. At these depths the threshold of most JFETs is outwith typical operating voltages (~5V).

Vertical pnp transistors are the only vertical bipolars offered by the EMF process. The p+ source/drain implant becomes the emitter implant. The p+ source-drain implant of $2x10^{15}$ atoms/cm^2 results in an emitter which will be less heavily doped than a traditional bipolar transistor. Once again this will reduce the emitter efficiency as the ratio of emitter/base doping concentrations becomes lower. The junction depth of the emitter will be around 1μm with the EMF process. This will result in an emitter with a high resistance when compared to those devices with emitter sheet resistances of around ~20 Ω/sq.

Thus the EMF process will be able to fabricate functional bipolar and JFET devices. However, the performances of these devices will be poor compared to traditional devices. The purpose of this project was to see if changes in the performance of these devices can be related to process variations. As a result, absolute performance is not a critical factor. Thus from a preliminary study it appeared to be feasible to fabricate bipolar and JFET devices using the EMF 5μm CMOS process.

4.3.3 Initial device simulation

Before designing test chips for this project, device simulation tools available to the EMF were utilized for dimension and electrical characterisation. These included SUPREM3, SUPRA and PISCES. All of these simulation packages came from the TMA software house.

SUPREM3 and SUPRA are physical device simulators, SUPREM3 is a 1-D simulator whereas SUPRA is a 2-D device simulator. Figure 4.2 shows a typical SUPREM3 profile of the EMF n-well with p+ source-drain implant. This was after 18hrs drive-in. Figure 4.3 shows a typical SUPREM3 input file required to generate this output. Figure 4.4 shows a SUPRA simulation of a vertical bipolar fabricated with the EMF CMOS process. This type of simulation gave details of side diffusions and helped establish minimum geometries for the test chip design. Figure 4.5 shows typical input files for SUPRA.

These simulation packages were used to determine junction depths for the parasitic transistors fabricated. Chapters 5 and 7 detail relationships between device performance and simulated junction depth.

4.4 Test Chip Design

Two test chips were designed for this project. They were designed using the GAELIC design suite. The design work was carried out on a TEKTRONIX 4014 terminal connected by modem to a PRIME computer in Oxford. This provided monochrome graphics on an electrostatically refreshed screen. Although CAD technology has advanced to reduce the design time associated with more complex structures, the design tools available for this project were well suited to the manipulation of rectangles and polygons. These were the basic building blocks for the test chip designs detailed later in this chapter.

To fabricate the type of parasitic structures described above certain of the design rules had to be violated. One of these was minimum spacing. Although the design set is ostensibly for 5μm design rules, the EMF fabrication facility had step and repeat capability with a GCA stepper. This allowed minimum spacings of 2μm to be included in some of the lateral bipolar designs.

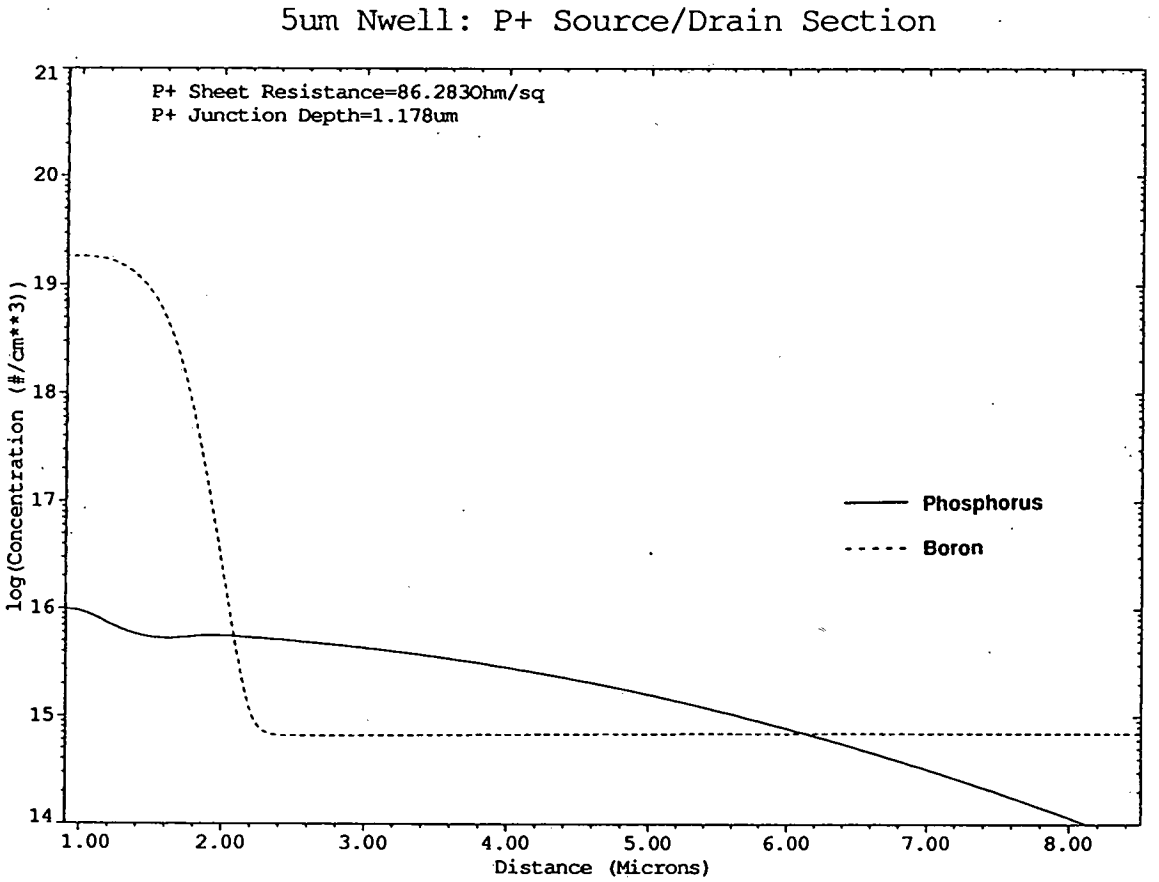


Figure 4.2. Sample SUPREM3 profile simulating the EMF 5μm process after p+ implant and well drive-in. This does not represent any of the devices fabricated.

```
*******************************************
***                SUPREM-3               ***
***          version 3C rev. 8636         ***
***  Copyright (C) 1983, 1984, 1985, 1986 ***
***  Technology Modeling Associates, Inc  ***
***            All Rights Reserved        ***
*******************************************

                30-DEC-90  13:08:54

   Statements input from file backend.dat

 1... comment this part to simulate the backend heat treatment

 2... initialize structure=fldoxg10

 3... comment anneal the implants
 4... diffusion temp=950 time=25 dryo2

 5... comment reflow pyro
 6... diffusion temp=1045 time=10 nitrogen
 7... diffusion temp=945 time=25 nitrogen

 8... comment second reflow
 9... diffusion temp=1050 time=10 nitrogen

10... plot active net left=0 right=3 dev=4010
11... plot boron active color=2 line=2 add
12... plot phosphor active color=3 line=3 add

13... print layer
14... savefile structure file=gatf10
```

Figure 4.3. Typical SUPREM3 input file used to generate the output shown in figure 4.2.

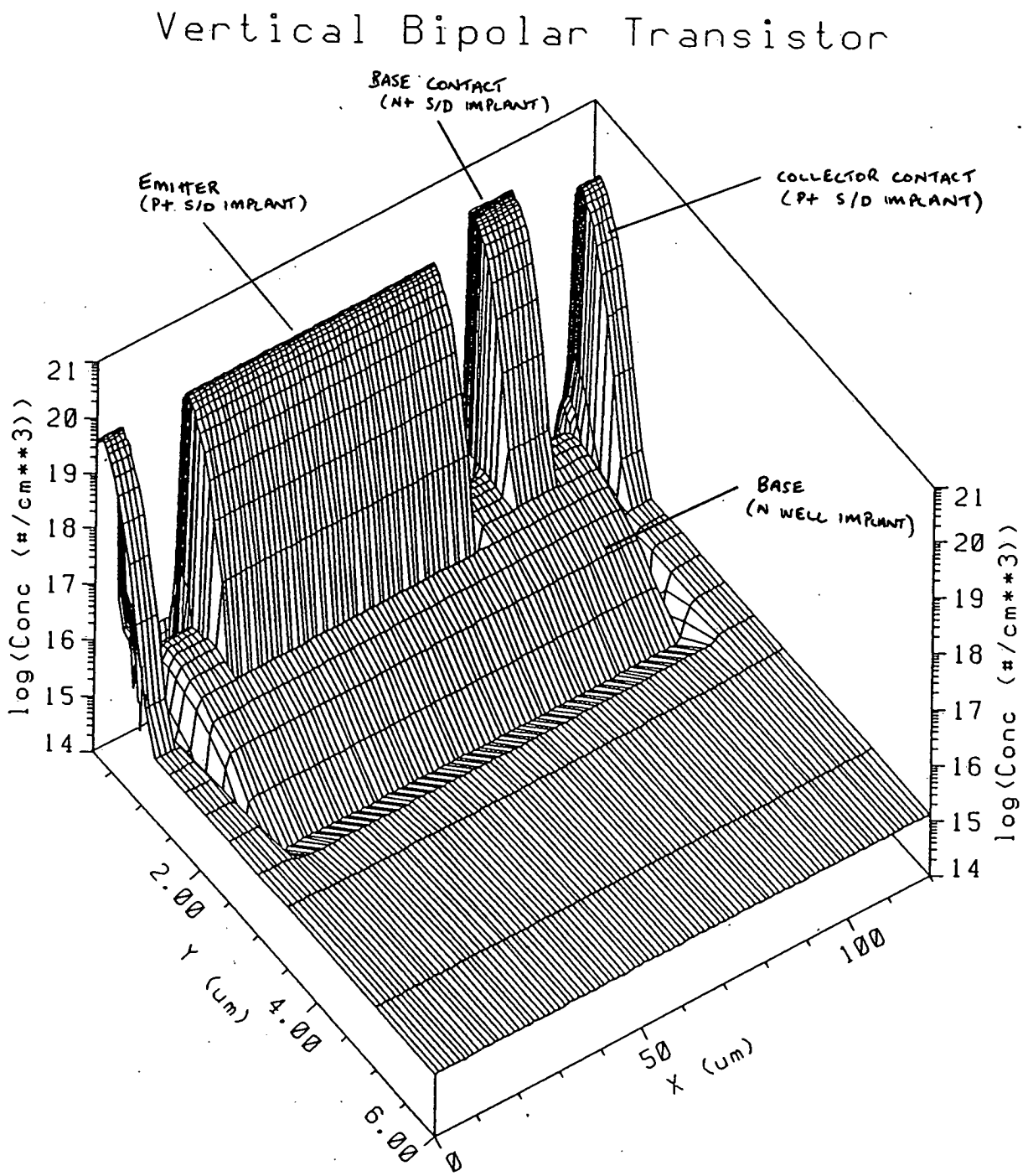


Figure 4.4. Sample SUPRA profile simulating the EMF 5µm process after p+ implant and well drive-in. This does not represent any of the devices fabricated.

```
title SUPRA Ldd structure source-drain implant
comment Numerical source/drain simulation

comment read in the structure
structure
load structure file=lddls
end

comment switch to numerical mode
numerical

coment define polysilicon gate

deposit polysilicon gate thick=.5
etch polysilicon x3=4.8 x4=5

comment implant the source/drain region
implant arsenic dose=5e13 energy = 90

plot.2d scale device=hp7550
end

 savefile all file=depictin
```

Figure 4.5. Typical SUPRA input file used to generate 2D simulations.

4.4.1 The Lateral Bipolars

Several types of lateral transistor were designed. These are shown in figures 4.6 a-d. Figure 4.6 a is the simplest. This design involves one p+ stripe as the emitter and one p+ stripe as the collector. The effective emitter area is proportional to the length of the stripe multiplied by the junction depth (for an ideal lateral). This is investigated in more detail in chapter 6. The base region is formed by the n-well implant. The base width is determined by the drawn spacing between the emitter and collector stripes. Contact to the base region is made through an n+ implant in the n-well. Arrays of these transistors were designed with drawn base widths of 2μm to 4.5μm in 0.5μm steps. Identical arrays were also designed for n-p-n transistors. These utilize the n+ implant direct into substrate for the emitter and collector. The silicon substrate forms the base regions. An p+ implanted substrate contact acts as the base contact.

Figure 4.6 b shows the same type of stripe lateral transistor. However, this device has a twin collector structure. The twin collector structure doubles the size of the effective emitter area and thus increases the gain of the device. Arrays of these transistors with the same drawn base width dimensions as those above were designed as both p-n-p and n-p-n devices.

Figure 4.6 c illustrates a lateral device designed with comb emitter-collector structures. This design maximises the effective emitter area for the lateral devices. However, the increased gain of these devices may be offset by the detrimental effect of having so many corners in the structure. The field from emitter to collector will by no means be constant along the length of the emitter. This will affect the emitter injection along the length of the transistor and perhaps produce some unexpected characteristics. Arrays of these devices were designed with varying drawn base widths.

Figure 4.6 d shows a lateral device with a totally enclosed circular emitter. This design increases the effective emitter area but does not introduce any of the variations in emitter-collector field present in the comb structure. These transistors were designed with both varying emitter-collector width and varying emitter radius. Although the prime design suite offered circular designs for the emitter, there was a

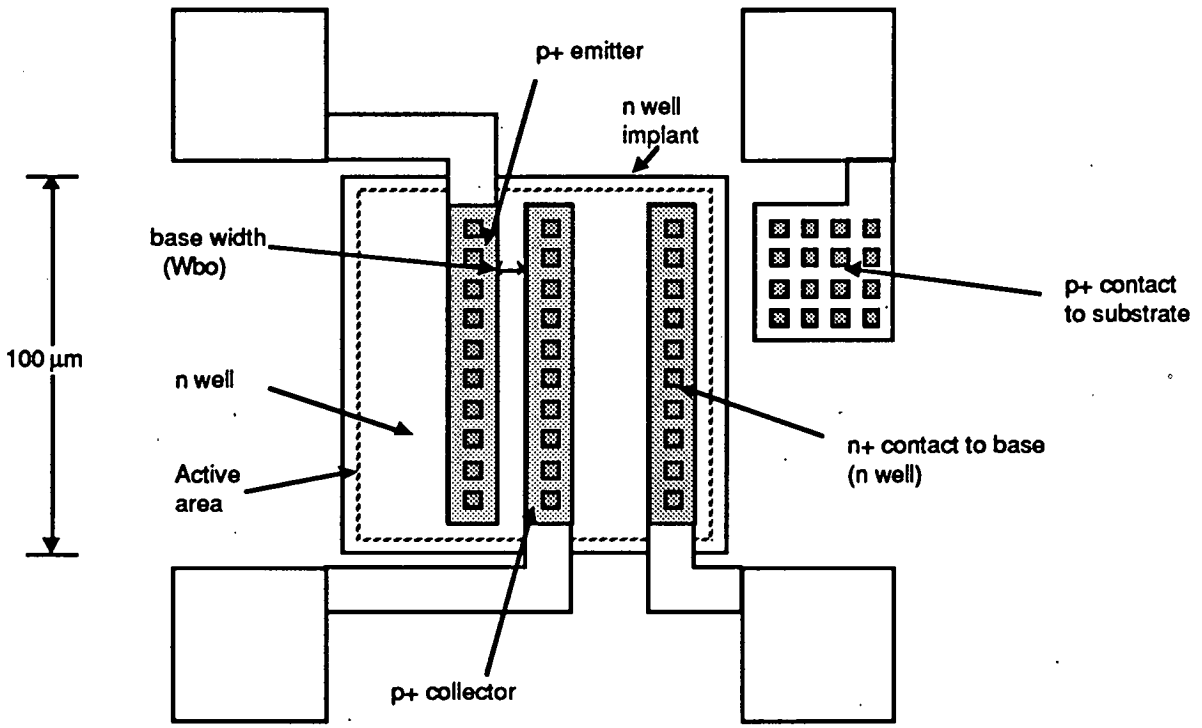


Figure 4.6a. Simple lateral transistor with single emitter and collector stripes

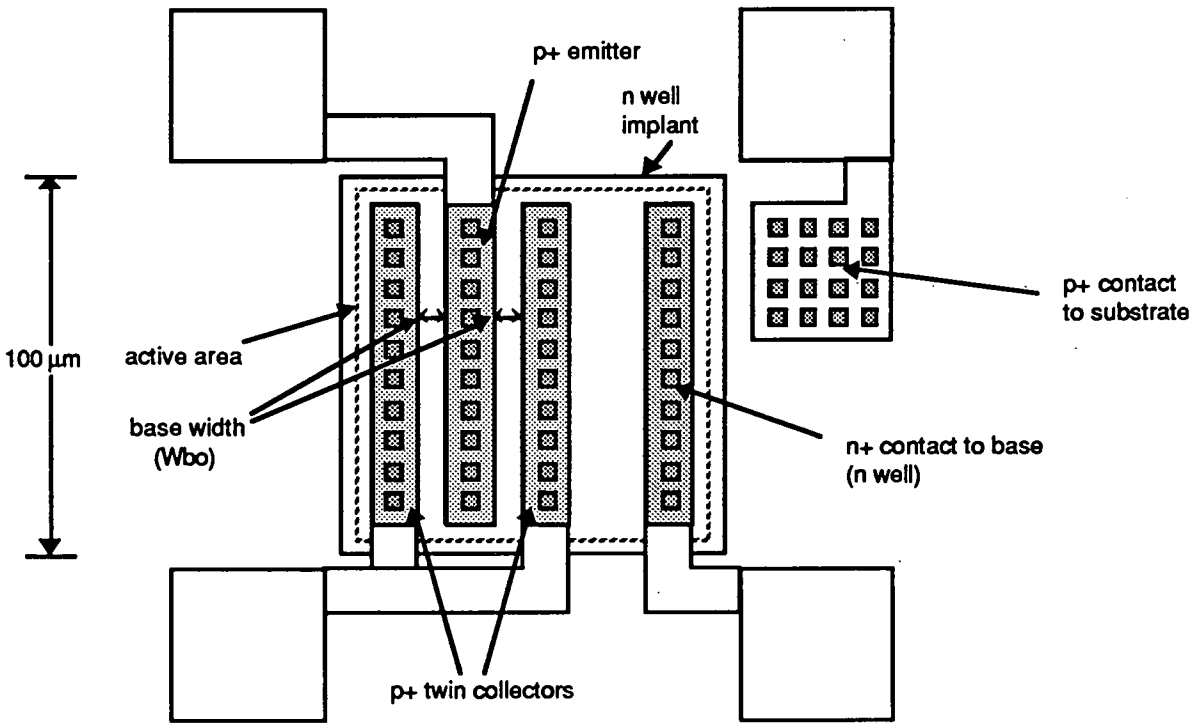


Figure 4.6b. Lateral bipolar with stripe structured emitter and twin bar collector.

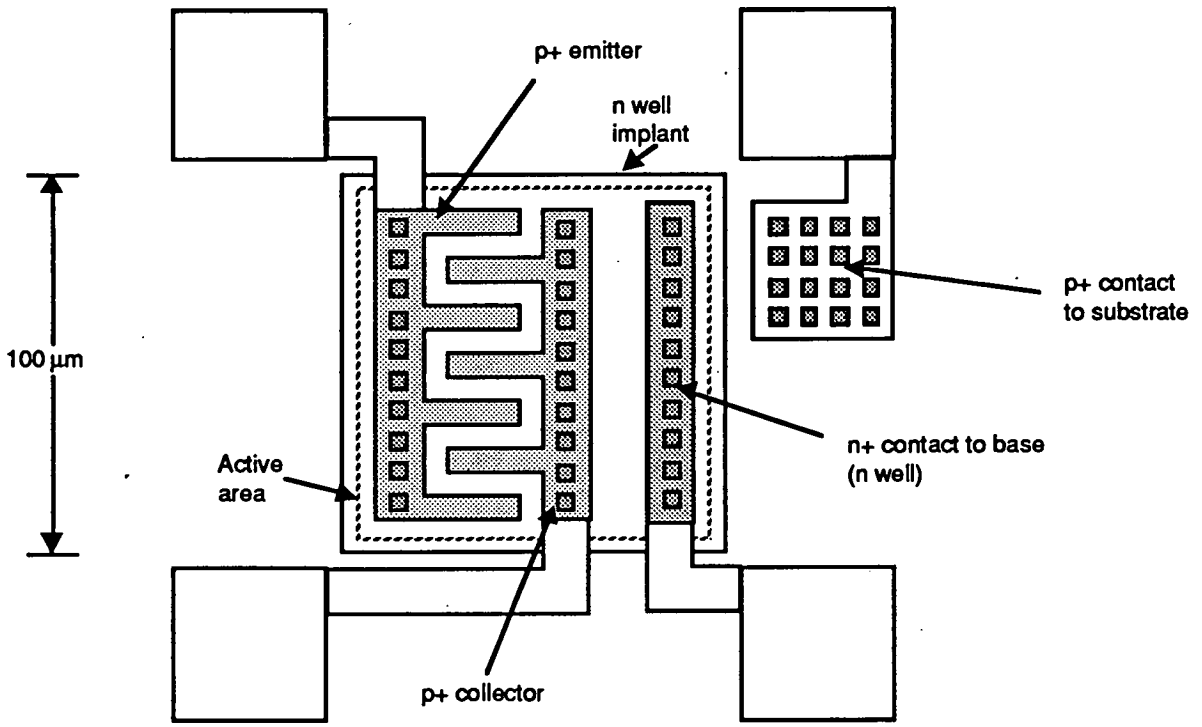


Figure 4.6c. Lateral bipolar with comb structured emitter and collector.

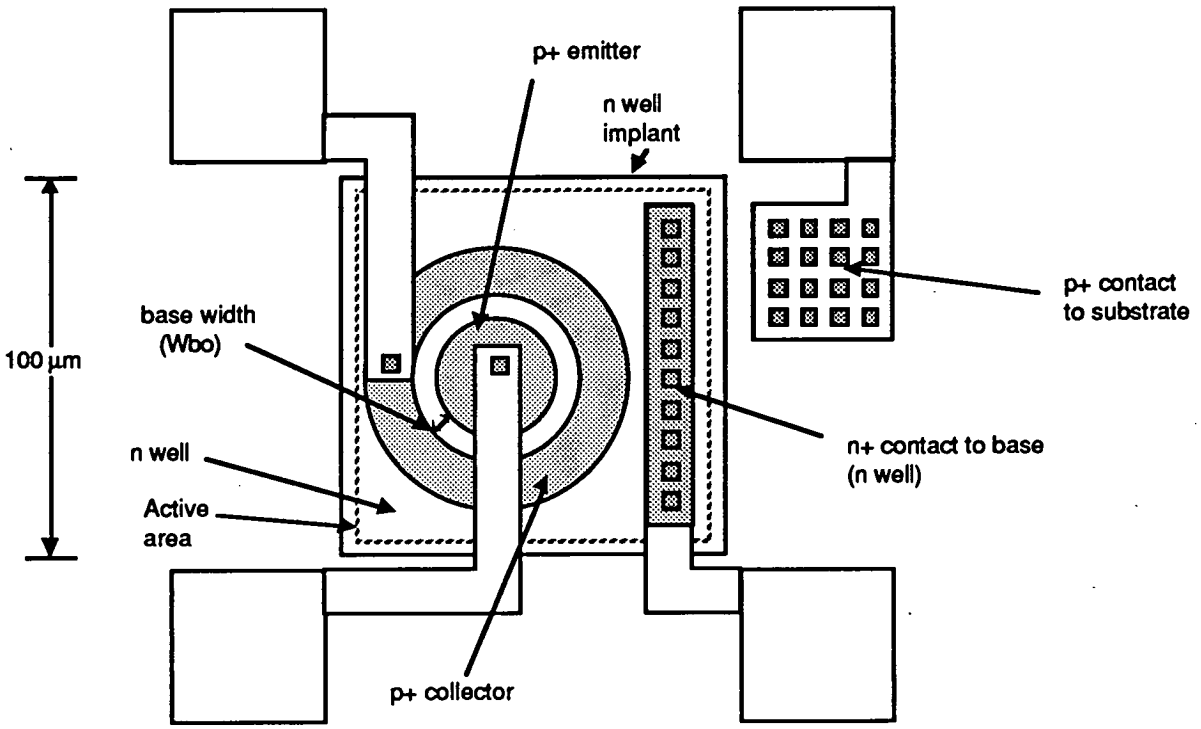


Figure 4.6d. Lateral bipolar with circular structured emitter and collector.

problem with the design of the collector. In order to surround the emitter, the collector had to be stitched together, ie a stripe had to be 'bent' around the emitter. This caused problems at the mask making stage, the consequences of which are dealt with in chapter 6.

4.4.2 The Verticals

Figure 4.7 a-c illustrates the basic design of the vertical bipolar devices. The emitter region is formed by a p+ implant. The base region is formed by the n-well implant. Base contact is made through an n+ contact to the n-well. The substrate forms the collector. Contact to the collector is made with a p+ contact to the substrate. Contact to the collector is also possible from the back of the wafer provided any oxide is removed. Figure 4.7a shows a fairly straightforward design with a rectangular emitter and a stripe base contact. Figure 4.7b shows the same design but with an alternative collector contact. Arrays of these devices were designed with 3 different emitter sizes. These were 70μm x 50μm, 50μm x 30μm and 40μm x 20μm.

Figure 4.7c shows a design which was created but never fabricated. This device uses the p+ implant as a base. The emitter is the n+ implant. This would 'push out' the base implant. The collector is the n-well implant. This device would be a n-p-n device with a very narrow, highly doped, base region. For the device to be feasible, the p+ and n+ source-drain implants would have to be reversed in order. Also great care would have to be taken not to punch the emitter junction through to the collector. The design was included for possible future evaluation and would be relevant to a twin well CMOS process.

4.4.3 The JFETs

Figure 4.8 illustrates the two types of JFET designed for this project. The gate of the JFET is formed by the p+ implant. The channel region is created with the n-well implant. Source-drain contacts are made on either side of the gate with the n+ implant. The difference between the designs is the overlap of the gate junction, (figure 8b). The gate is overlapped in the active area to make contact with the substrate and effectively shorts the gate

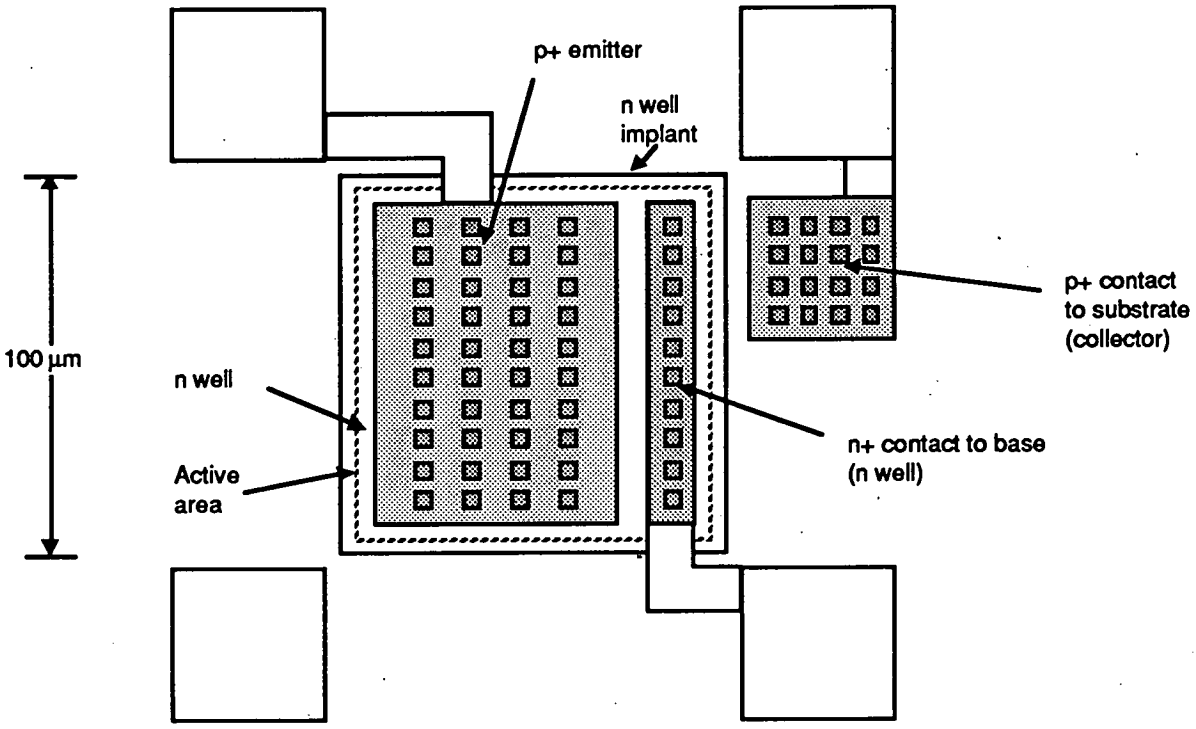


Figure 4.7a. Typical vertical bipolar design showing use of CMOS process implants.

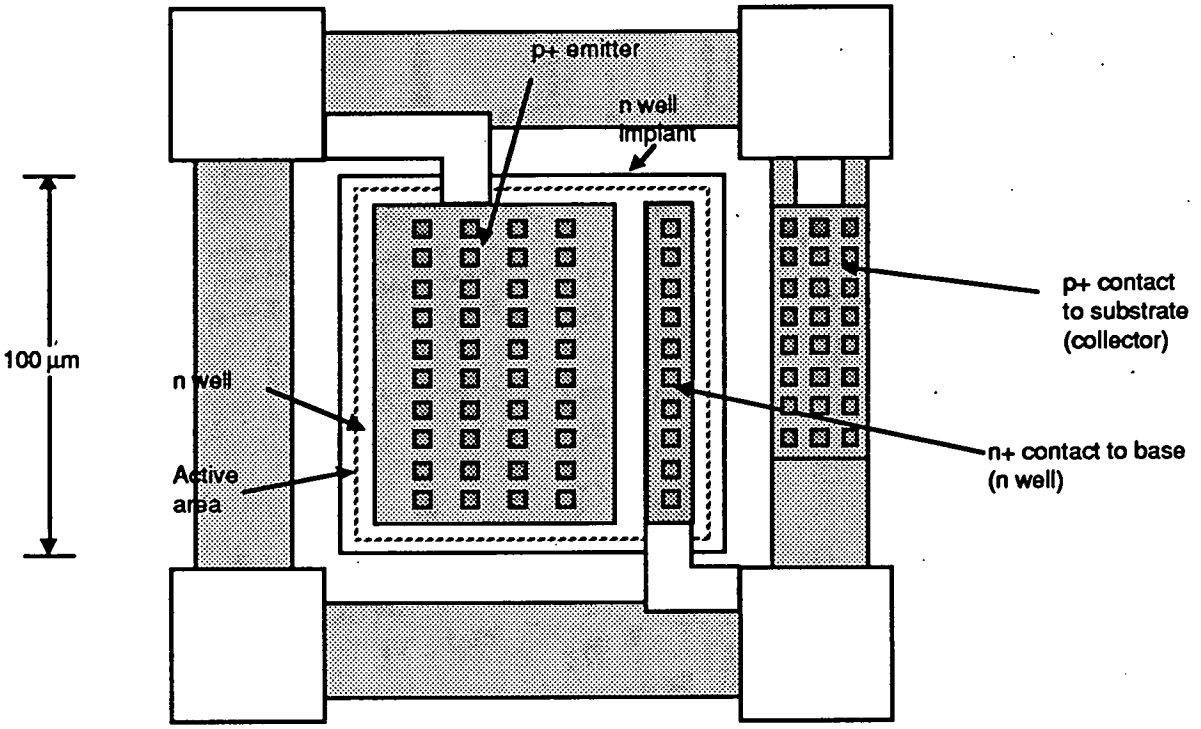


Figure 4.7b. Typical vertical bipolar design with alternate collector structure. In this device the collector completely encloses the implanted base region.

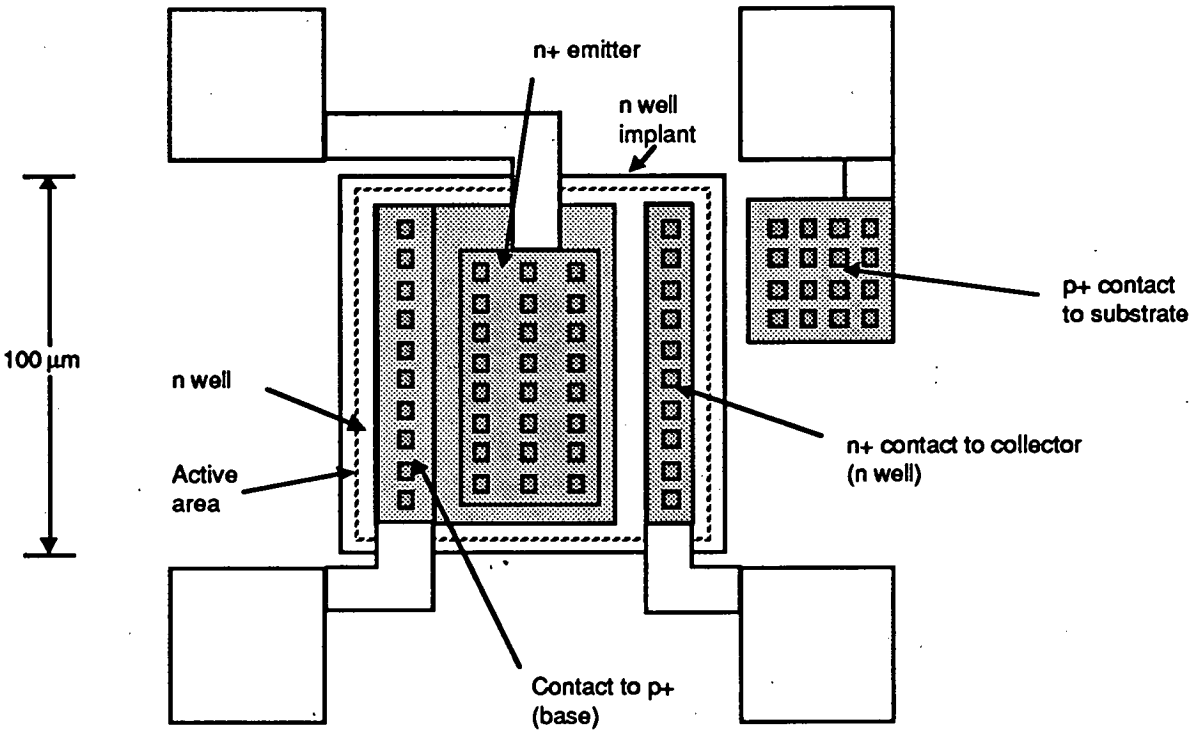


Figure 4.7c. Narrow base bipolar design. This transistor would be feasible if the order of the p+ and n+ implants were reversed.

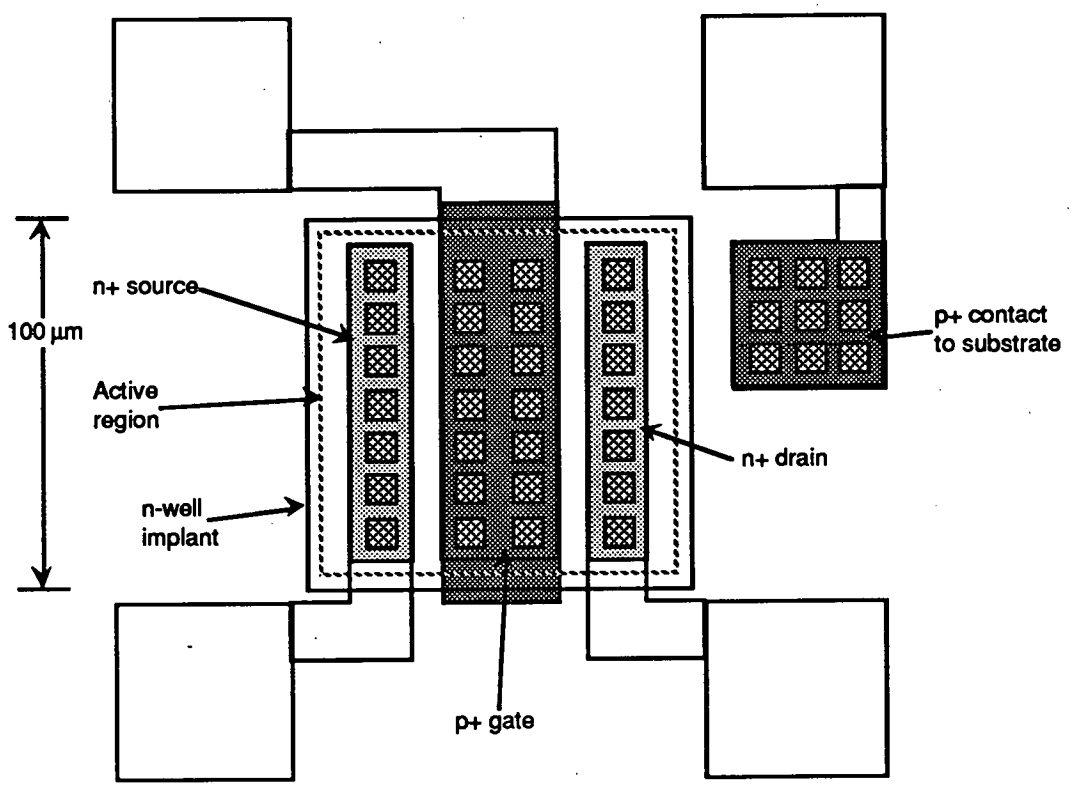


Figure 4.8a. Schematic of JFET design with non overlapping gate.

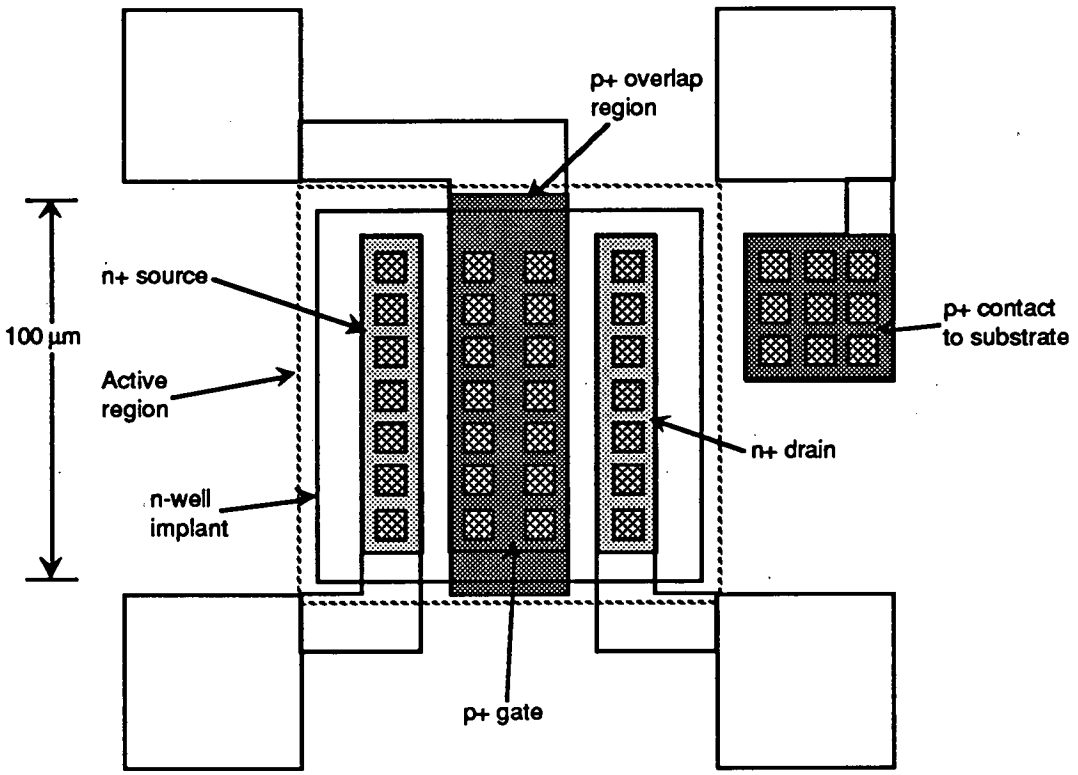


Figure 4.8b. Schematic of JFET design with overlapping gate. In this design the p+ gate is in direct contact with the p- substrate. This feature effectively clamps the substrate to the gate potential.

and the substrate together. This has repercussions on the source-drain leakage of the device and the threshold V_T. Both these effects are discussed in more detail in chapter 5. Arrays of these devices were created by varying the gate width from 25μm to 30μm in 1.25μm steps.

4.4.5 The MOSFETs

The aim of this project was to investigate the feasibility of using 'parasitic' devices for CMOS process control. In order to compare any techniques developed, MOS transistors had to be included in the test chip design. Figure 4.9 illustrates a typical MOS design.An array of these devices was designed. These had dimensions W/L = 25/25, 25/10, 25/5, 25/3, 25/2, 2/25, 3/25, 5/25, and 10/25 (all dimensions are in μm).

4.4.6 The Layout

It was decided that a 2xn pad layout should be adopted for both test chips. This type of layout offers many advantages. All structures can be easily probed and require only one probe card. Figure 4.10 shows the pad layout for one of the test chips. As discussed in chapter 3, the characterisation hardware was provided by a Hewlett Packard 4145. No switching matrix was available for the set-up. This limits the testing to one device at a time as there are only 4 stimulation/measurement units available and these cannot be switched. As a result a 4 pin probe card was used to test the devices. Both test chip designs were incorporated into one mask set. The masks were made in-house at the EMF.

4.5 Well Depth Process Split

To investigate the possible use of parasitic devices for monitoring n-well junction depth, a process split was made on a batch of 10 wafers. An n-well junction depth split was chosen because it was felt that the operation of both the vertical bipolar transistors and the JFET transistors would be sensitive to this parameter. With the trend in modern CMOS towards much shallower well depths (chapter 1) this becomes a critical factor in circuit operation. Both latch-up characteristics and leakage effects are much influenced by well junction depth. Conventional junction measurement techniques are nearly all

destructive (chapter 2). It was hoped from this split that an electrical method of well junction depth monitoring could be achieved. Details of the split and device results are given in chapters 5 and 7.

4.6 Other Applications

This chapter has given the details and rationale behind the design of two test chips for this project. Lateral and vertical bipolar designs established here were adapted for use in COMETT project 643D. This project produced a VLSI teaching chip for use in universities and industry.

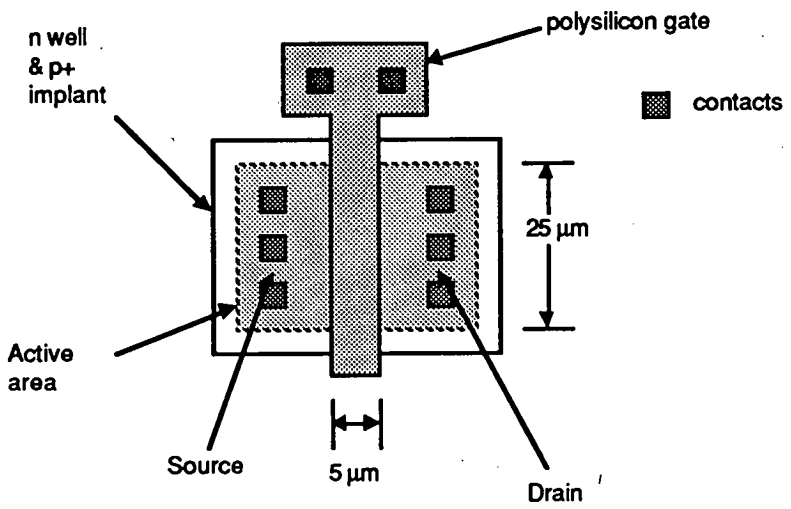


Figure 4.9. Schematic of typical MOS design.

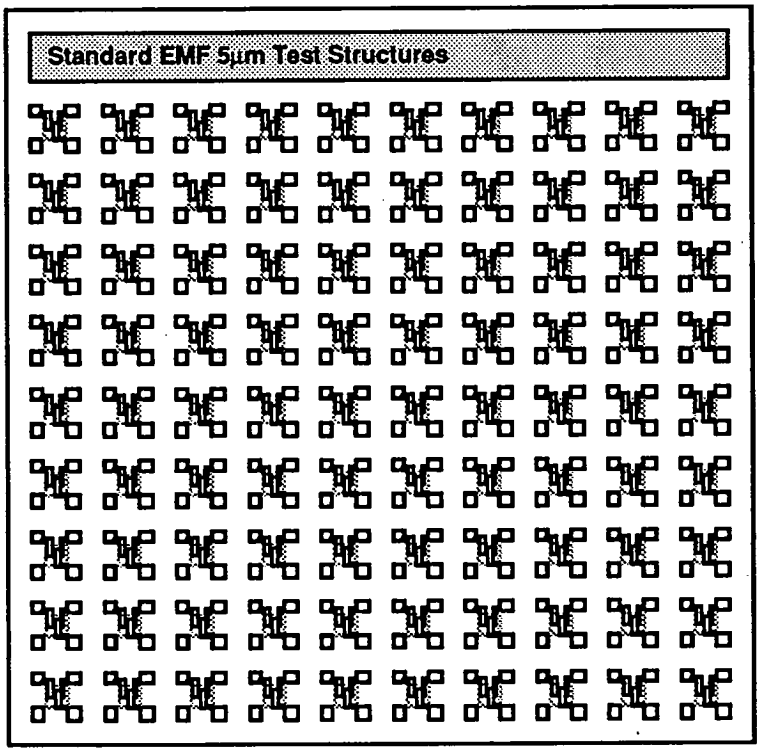


Figure 4.10. 2xn pad layout chosen for the two test chips designed. This facilitated simple probing with 1 probe card.

References

[1] W.M. Webster, *Proc. IRE*, 42, p.914, 1954.

[2] R. S. Muller and T. I . Kamins,"Device Electronics for Integrated Circuits,"*Wiley 2nd edition*, p303, 1986.

[3] R. S. Muller and T. I . Kamins,"Device Electronics for Integrated Circuits,"*Wiley 2nd edition*, p327, 1986.

Chapter 5

The use of parasitic JFETs to monitor Well depth

5.1 Introduction

JFET transistors can be fabricated in a standard n-well CMOS process and no additional masks or implants are required. The electrical characteristics of these parasitic JFETs depend on n-well depth and doping concentrations as well as source/drain junction depths and concentration profiles. Arrays of parasitic JFETs were designed and fabricated using the EMF 5μm n-well CMOS process (chapter 4). Electrical characterisation of parasitic JFET transistors provide parameters that can be used for in-line monitoring of implant profiles and related diffusion cycles. Figure 5.1 shows a schematic diagram of a typical parasitic JFET fabricated using a n-well CMOS process.

5.2 JFET electrical theory

The depletion layer width of a reverse biased pn junction can be varied by modulating the size of the reverse bias voltage. A JFET uses this mechanism to control the current through a region bounded by one

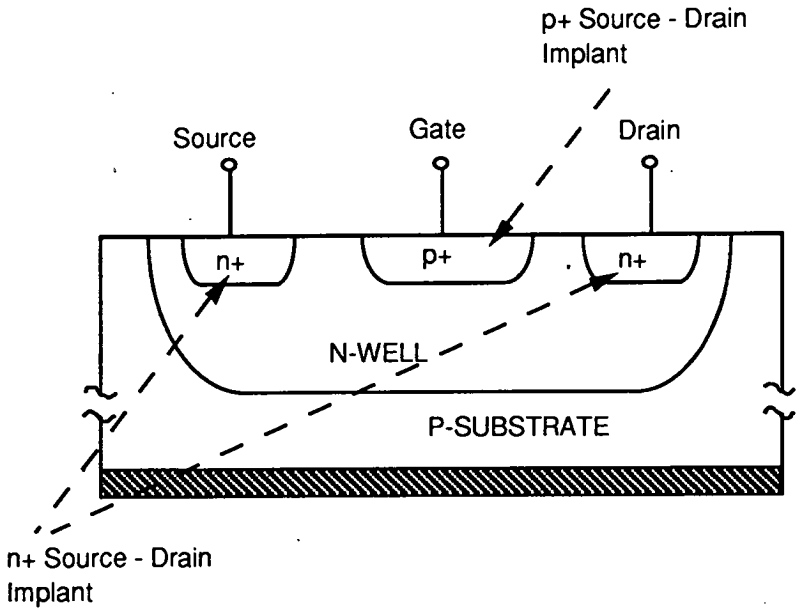


Figure 5.1 Schematic of a typical parasitic JFET fabricated in a standard CMOS process.

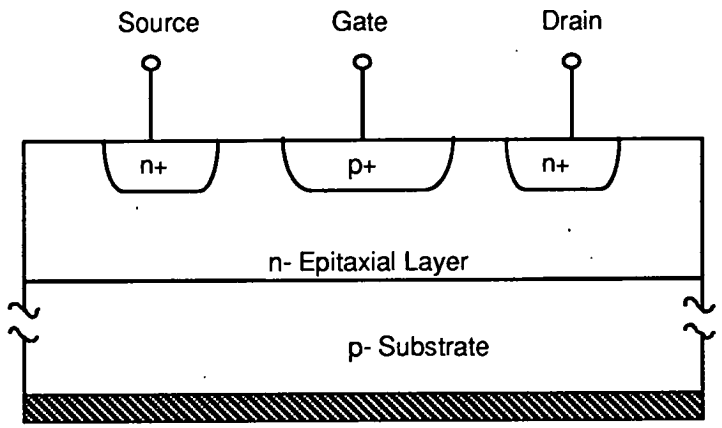


Figure 5.2 JFET fabricated using epitaxy to provide a homogeneous lightly doped channel.

or more pn junctions. Very little power is consumed at the pn junction itself. However a much greater power can be controlled and delivered through the current in the device channel. The JFET can therefore be used as a power transistor. A typical JFET is shown in Figure 5.2. It consists of a lightly doped n-type layer on top of a p-type substrate. The n-type region is usually an epitaxially grown layer. This provides a uniform and well-controlled dopant concentration. This layer could also be created by ion implantation. Source and drain connections are added by diffusion of n+ electrodes at either side of the channel. The source is the electrode that supplies majority carriers to the channel. Consequently, conventional current flows from drain to source in an n-channel JFET.

The pn junction above the channel serves as the control element. When a reverse bias is applied to it, this serves as a gate. The channel is defined from the depletion layer edge of the pn junction at the gate to the depletion layer edge of the substrate pn junction. The substrate is usually grounded and a positive bias applied to the drain. If a negative bias is applied to the gate, the depletion layer around the gate region increases which in turn narrows the channel and increases it's resistance. The increase in resistance restricts the current flowing from drain to source. Thus a signal applied to the gate can control the current flowing through the channel.

5.3 Device Analysis

Many text books [1,2,3] contain a detailed description of the device physics involved in JFET operation so only a simple analysis is given here following that of Muller and Kamins [4]. We will consider the situation where a small bias V_D is applied to the drain electrode while the source is grounded. Under this condition the gate-channel bias and therefore the width of the gate depletion region is uniform along the length of the entire channel. The voltage at the gate is V_G. An expanded view of the channel is shown in Figure 5.3. We assume a one-dimensional structure with a gate length L between the source and drain regions and a width W perpendicular to the plane of the paper (usually $W >> L$). Drain current flows along the dimension y. We assume a one-sided step junction at the gate with a dopant density N_a in the p+region much greater than N_d in the channel. The distance

between the p-type gate and the substrate is t. The thickness of the gate depletion region in the n-type channel is x_d and the neutral channel width is x_w. One major assumption is made: the depletion region at the substrate extends primarily into the substrate so that $x_w \sim (t - x_d)$. This is approximately true in practice.

The channel resistance is given by :

$$R = \frac{\rho L}{x_w W} \tag{5.1}$$

where $\rho = (q\mu_n N_d)^{-1}$ is the resistivity of the channel. Hence the drain current is

$$I_D = \frac{V_D}{R} = \frac{W}{L}\left(q\mu_n N_d x_w V_D\right) . \tag{5.2}$$

A dependence on gate voltage is incorporated into equation 5.2 by expressing x_w as $t - x_d$ where x_d is

$$x_d = \left[\frac{2\varepsilon_s}{qN_d}(\phi_i - V_G)\right]^{\frac{1}{2}} . \tag{5.3}$$

$$I_D = \frac{W}{L}\left(q\mu_n N_d t\right)\left\{1 - \left[\frac{2\varepsilon_S}{qN_d t^2}\left(\phi_i - V_G\right)\right]^{\frac{1}{2}}\right\} V_D \tag{5.4}$$

The multiplying factor at the front of the brackets is known as the conductance G_o of the undepleted channel region. Equation 5.4 becomes

$$I_D = G_o\left\{1 - \left[\frac{2\varepsilon_S}{qN_d t^2}\left(\phi_i - V_G\right)\right]^{\frac{1}{2}}\right\} V_D . \tag{5.5}$$

From Equation 5.5 we see that this device has a linear relationship between V_D and I_d for a given gate voltage. This is a consequence of having assumed a small applied drain voltage. The square root dependence on gate voltage in Equation 5.5 is a result of the abrupt gate-

channel junction. (Abrupt channel theory was outlined in chapter 2). Equation 5.5 will give maximum current for zero applied gate voltage and decreases as $|V_G|$ increases. Equation 5.5 predicts zero current when the gate voltage is large enough to deplete the entire channel region.

We can also remove the restriction of small applied drain voltage and consider the case of arbitrary V_D and V_G values (the only restriction is that the gate must always remain reverse biased). With a fixed V_D, the potential difference between the gate and channel is a function of position y. Thus the depletion width and resulting channel cross-section is a function of position y. The voltage across the depletion region is higher near the drain than the source in this n-channel device. Therefore, the depletion region is wider near the drain as shown in Figure 5.4. To derive an expression for the drain current we use what has been called the gradual channel approximation. This approximation assumes that the channel and depletion layer widths vary slowly from source to drain so that the depletion region is influenced only by fields in the vertical dimension and not by fields extending from drain to source, i.e., the field in the y direction is much less than that in the x direction in the depletion analysis. Thus the width of the depletion region can be found by a one dimensional analysis. Using this approximation, we can write an expression for the increment of potential across a small section of the channel of length dy at y as

$$d\phi = I_D dR = \frac{I_D dy}{Wq\mu_n N_d(t - x_d)} . \tag{5.6}$$

The width x_d of the depletion region is now controlled by the voltage $\phi_i - V_G + \phi(y)$ where $\phi(y)$ is the potential in the channel at point y and ϕ_i is the built-in potential, so that

$$x_d = \left[\frac{2\varepsilon_S}{qN_d}\left(\phi_i - V_G + \phi(y)\right)\right]^{\frac{1}{2}} . \tag{5.7}$$

By substituting this expression into Equation 5.6 and integrating over the channel length from source to drain we obtain the current-voltage relationship for the JFET.

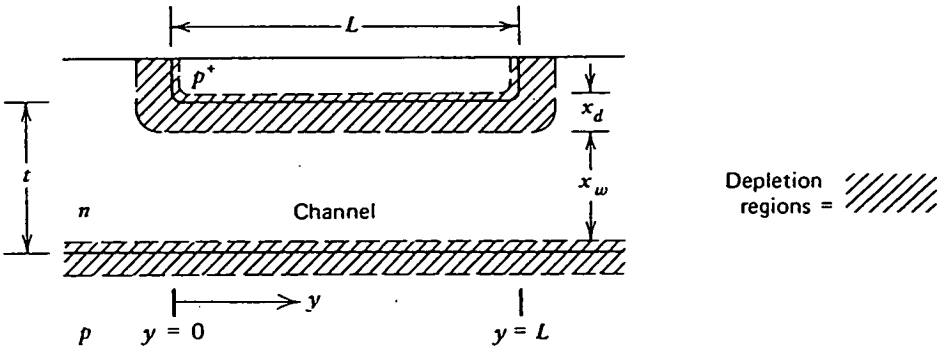


Figure 5.3 Channel region of a JFET with gate length L, showing depletion regions. Typical dimensions might be L=8μm and t = 1μm.[4]

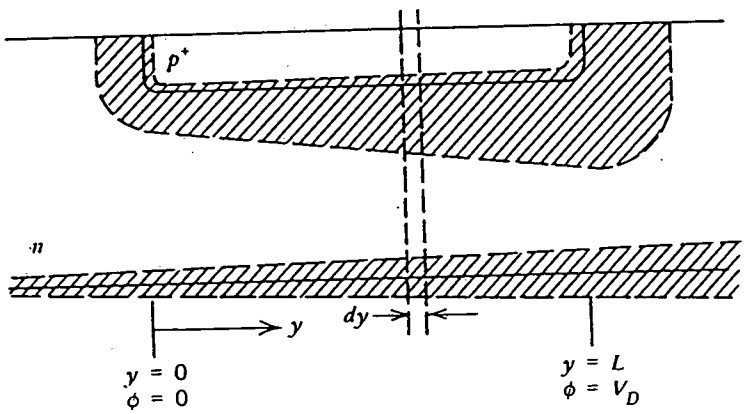


Figure 5.4 Channel region of a JFET showing variation of the width of depletion regions along the channel when the drain voltage is significantly higher than the source voltage.[4]

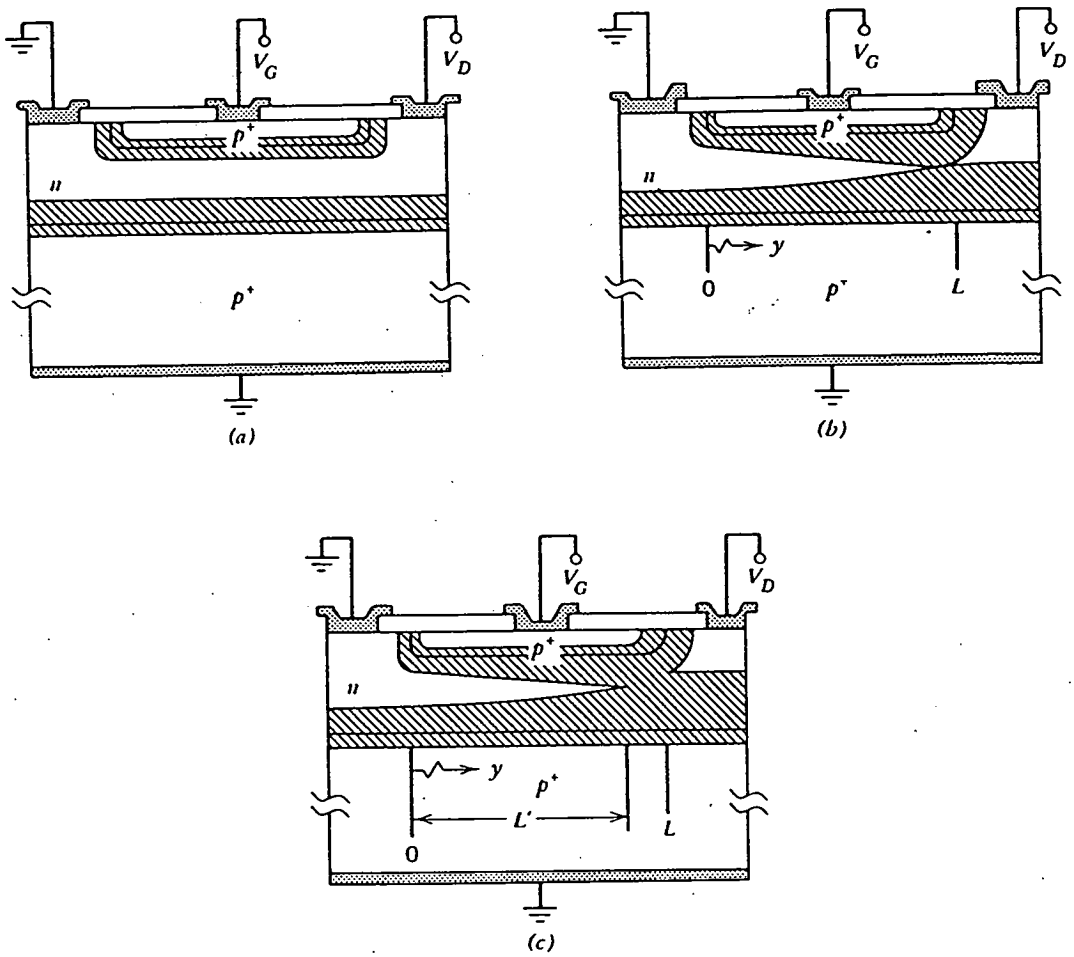


Figure 5.5 Depletion layer behaviour in a JFET. (a) Small drain voltage, channel is approximately at equipotential with constant resistance along it's length. The dimensions of the depletion regions are uniform. (b) $V_D = V_{Dsat}$, the depletion regions on both sides of the channel meet at the pinch off point (at $y=L$). (c) $V_D > V_{Dsat}$, the pinch off point at ($y = L'$) moves slightly closer to the source.[4]

$$\frac{I_D \int_0^L dy}{Wq\mu_n N_d} = \int_0^{V_D} \left\{ t - \left[\frac{2\varepsilon_S}{qN_d} \left(\phi_i - V_G + \phi(y) \right)^{\frac{1}{2}} \right] \right\} d\phi \qquad (5.8)$$

After integrating and re-arranging

$$I_D = G_O \left\{ V_D - \frac{2}{3} \left(\frac{2\varepsilon_S}{qN_d t^2} \right)^{\frac{1}{2}} \left[\left(\phi_i - V_G + V_D \right)^{\frac{3}{2}} - \left(\phi_i - V_G \right)^{\frac{3}{2}} \right] \right\}. \qquad (5.9)$$

At low drain voltages Equation 5.9 reduces to that of Equation 5.5 and the current increases linearly with drain voltage. At higher drain voltages, however, the current increases more gradually. As the drain voltage increases, Equation 5.9 shows the current in the channel increasing to a maximum and then decreasing. The maximum corresponds to the limit of validity of this one-dimensional analysis. From Figure 5.5a we can see that as the drain voltage increases the width of the conducting channel near the drain decreases until the channel is completely depleted in this region (Figure 5.5b). In this situation Equation 5.6 cannot be resolved ($x_d \rightarrow t$). This analysis is therefore only valid for drain voltages below the pinch off point of the channel. Although the channel is pinched off, current continues to flow because there is no barrier to the transfer of electrons travelling down the channel toward the drain. As they arrive at the edge of the pinched-off zone they are pulled across it by the field from the drain to the source. Any increased drain voltage is dropped across the depleted high field region near the drain and the depleted zone moves toward the source. If this movement is treated as slight then the current can be treated as constant (saturated). This bias condition is referred to as saturation. The drain voltage at which the channel is entirely depleted near the drain electrode is found from Equation 5.7 to be

$$V_{Dsat} = \frac{qN_d t^2}{2\varepsilon_S} - \left(\phi_i - V_G \right) \qquad (5.10)$$

and the corresponding drain current is

$$I_{Dsat} = G_O \left\{ \frac{qN_d t^2}{6\varepsilon_S} \cdot \left(\phi_i - V_G\right) \left[1 - \frac{2}{3} \left[\frac{2\varepsilon_S \left(\phi_i - V_G\right)}{qN_d t^2} \right]^{\frac{1}{2}} \right] \right\} . \quad (5.11)$$

From this analysis, the drain current-drain voltage characteristics may be divided into three regions (Figure 5.6)

(1) The linear region at low drain voltages.

(2) A region with less than linear increase of current with drain voltage.

(3) A saturation region where the current remains relatively constant as the drain voltage is further increased.

As expected from the physics of the device, Equation 5.11 predicts the current to be a maximum for zero gate bias and to decrease as negative gate voltage is applied. As the gate voltage becomes more negative, the drain saturation voltage and corresponding current decrease so that a family of curves may be generated (Figure 5.6), each curve showing the drain current versus drain voltage characteristic for a particular value of gate voltage. At a sufficiently negative value of gate voltage, the saturation drain current becomes zero. This turn off voltage V_T is found from Equation 5.11 to be

$$V_T = \phi_i - \frac{qN_d t^2}{2\varepsilon_S} . \quad (5.12)$$

The drain current actually increases slightly as the drain voltage is increased beyond V_{Dsat} because the end point for the integration in Equation 5.8 now becomes L' rather than L, where L' is the point at which the channel becomes completely depleted [$\phi(L') = V_{Dsat}$](Figure 5.5). For $V_D > V_{Dsat}$ the expression for I_{Dsat} in Equation 5.11 is multiplied

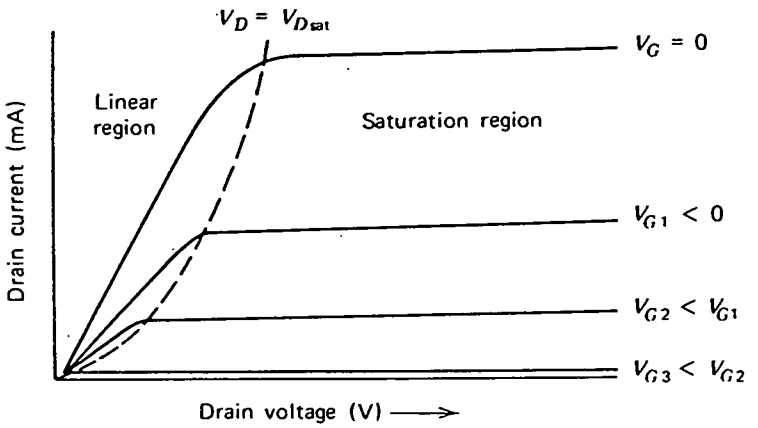


Figure 5.6 Output characteristics (drain-current, drain-voltage) characteristics of a JFET as a function of the gate voltage. Operation in the "linear" region (to the left of the $V_D = V_{Dsat}$ curve) corresponds to Figure 5.5a. In saturation (to the right of the $V_D = V_{Dsat}$ curve) Figure 5.5c applies.[4]

by the ratio L/L' (which is greater than unity). Although most JFETs exhibit characteristics with well-defined saturation regions shown in Figure 5.6, significant departures can be seen in devices with short channel lengths. In a device with a donor density of 10^{16} cm^{-3} in the channel region, an excess of 5V beyond V_{Dsat} depletes an additional 1.0 μm. Thus, for a typical channel length of 8μm, the ratio $L/L' \approx \frac{8}{7}$, and the deviation from simple theory is not severe. For channel lengths of the order of 2μm, however, important deviations occur. Channel length modulation for the MOS device was discussed along with other short channel effects in chapters 2 and 3.

5.4 JFET sensitivity to n-well depth

It is clear from the above analysis that the electrical characteristics of a JFET are very dependent on the channel region of the device. Thus the doping profile and junction depth of the channel region are perhaps the most important factors in the device operation. If a JFET is fabricated using an n-well CMOS process then the channel region will be formed by the n-well implant and drive-in sequence, as shown in figure 5.1. It follows that the characteristics of such a parasitic device will be sensitive to n-well processing. For this experiment, arrays of parasitic JFETs were designed and fabricated using the EMF 5μm process. The processing of these devices was subject to a n-well drive-in split. This split is detailed in chapter 4. The electrical characteristics of the devices were investigated using TECAP.

5.4.1 Characterisation model

The JFET model which is provided in TECAP is the single level model which is used in SPICE. This model relates the drain current to the square of the gate-source voltage. In addition, second order effects such as junction ohmic resistances, nonlinear junction capacitances, and channel length modulation are included.

The n-channel JFET model is shown in Figure 5.7. In the case of the p-channel JFET the polarity of node voltages V_{gs} and V_{ds} are reversed and the transistor is handled as an n-channel device The following equations are given for an n-channel device. The three regions

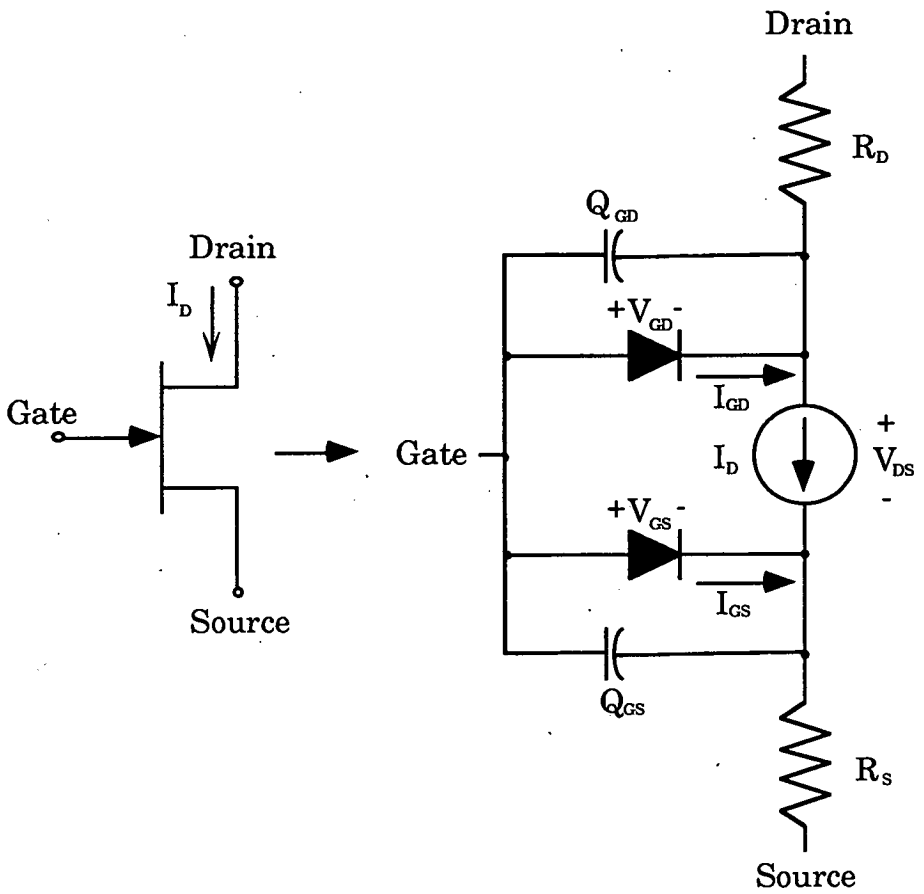


Figure 5.7. n-channel JFET model.

of operation discussed above are modeled using V_{gs}, the gate source potential and V_{ds}, the drain source potential. The equations are shown below

$$I_d = 0 \text{ for } V_{gs} < V_{to} \qquad \text{(cutoff)} \qquad (5.13)$$

$$I_d = \beta(V_{gs} - V_{to})^2(1 + \lambda V_{ds}) \qquad \text{for } 0 < (V_{gs} - V_{to}) < V_{ds}$$

$$\text{(saturation)} \qquad (5.14)$$

$$I_d = \beta V_{ds}\left[2\,(V_{gs} - V_{to}) - V_{ds}\right](1 + \lambda V_{ds}) \qquad \text{for } V_{gs} - V_{to} > V_{ds}$$

$$\text{(linear)}. \qquad (5.15)$$

Equation 5.14 is obtained by substituting the expression for ϕ_i in equation 5.12 into equation 5.9 and using the Taylor series expansion around the point $V_G = V_T$ [5]. β is the transconductance of the JFET in the saturation or ohmic region of operation.

$$\beta = \frac{W\mu\varepsilon_s}{tL} \qquad (5.16)$$

The relation shown in equation 5.14 is similar to the enhancement-mode MOSFET equation (chapter 2) with the channel depth t replacing the insulator thickness. In a graph of I_d versus V_{ds}, in the linear region of operation, β can be obtained from the slope of the curve. Equation 5.15 is obtained by performing the same substitution in equation 5.11.

λ is the channel length modulation factor used to represent the finite output conductance of the JFET in the saturation region. This is due to the second order effect of V_{ds} on the effective channel length (discussed above).

V_{to} is the zero bias threshold voltage used to model the gate turn-on voltage of the JFET. The convention in SPICE/TECAP is that V_{to} is always negative for both p and n channel devices. (In an n-channel JFET, V_{to} = -2 implies that threshold voltage is -2 volts. In a p-channel JFET, V_{to} = -2 implies that threshold voltage is +2 volts.)

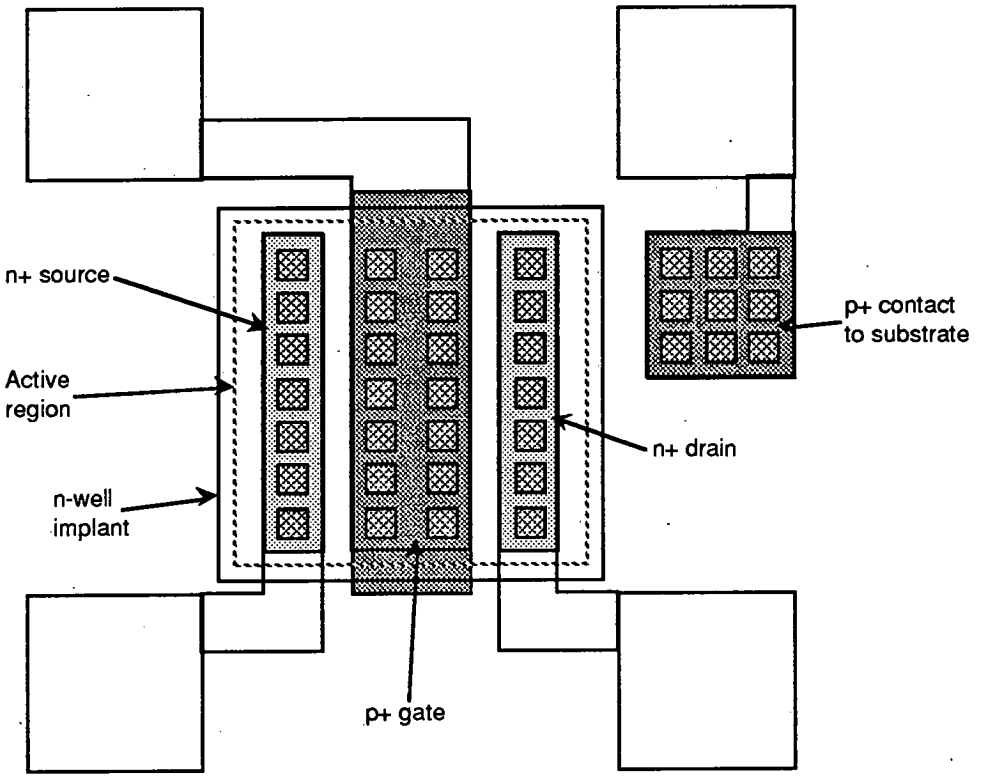


Figure 5.8a. Schematic diagram of typical JFET designed with non-overlapped gate. The active area of the device is enclosed by the n-well implant.

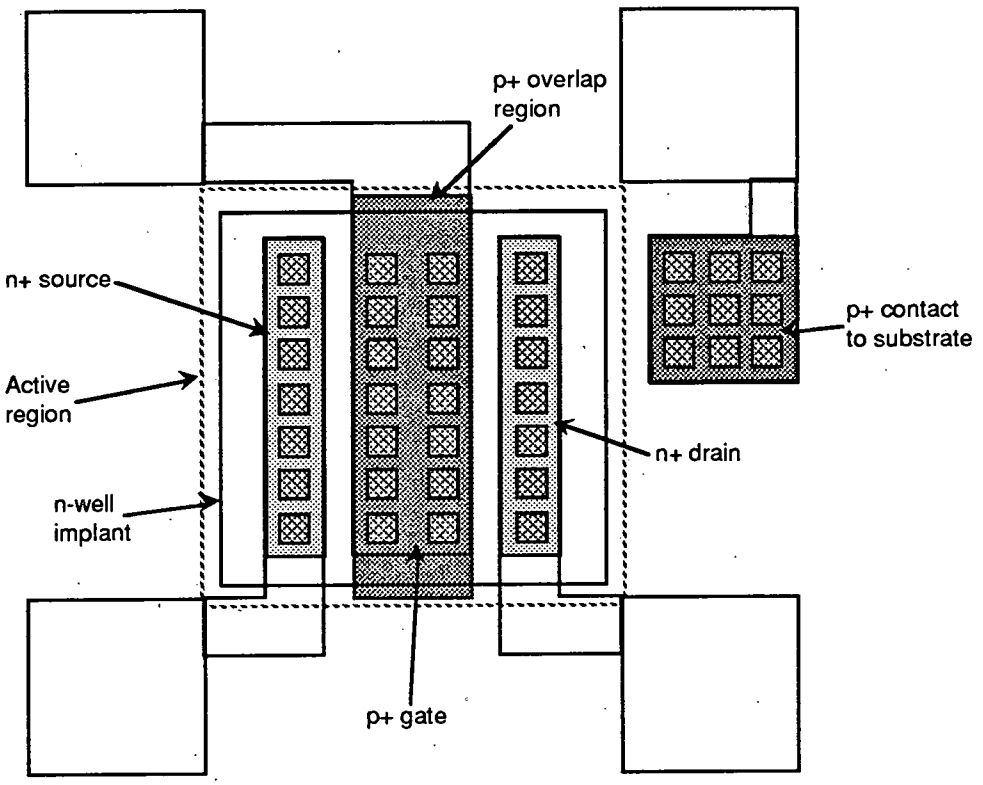


Figure 5.8b Schematic diagram of typical JFET designed with overlapped gate. The active area of the device extends beyond the n-well implant. This allows the p+ gate implant to make contact with the p-substrate.

5.4.2 Initial characterisation

Two types of parasitic JFET were initially investigated, Figure 5.8 shows a schematic of both structures. In Figure 5.8a the gate of the device is enclosed by the channel region, whereas, the device shown in Figure 5.8b has the gate region extending over the channel region into the p substrate. This means that the substrate in the second device is effectively clamped to the gate potential.

5.4.3 JFET Threshold Measurement.

JFET thresholds were measured using linear extrapolation of the I_D vs V_G curve. Threshold voltages were also obtained through the TECAP device characterisation. Although it is assumed some sort of linear interpolation is used in the TECAP program to calculate thresholds, for proprietary reasons the exact details of the method used is not made clear in the documentation. Thresholds were calculated for each of the devices in all parts of the split, Figures 9a-j show examples of each of these. From equation 5.12 the threshold voltage of a JFET is proportional to the square of the channel width (in a 1-d analysis), i.e.,

$$V_T \propto t^2 . \qquad (\text{for } V_T \gg \phi_i) \qquad (5.17)$$

From equation 5.17 it was expected that the thresholds obtained from the fabricated parasitic JFETs would be proportional to the depth of the implanted n-well. As discussed above the n-well forms the channel of the parasitic JFET, therefore t , the channel width, will be equal to the n-well depth minus x_j the gate junction depth.

5.4.4 Junction Leakage currents.

From Figures 9a-j we can see a marked increase in the leakage current I_s for the non-overlapped gate transistor when compared to the overlapped gate transistor. I_s arises from source-gate and drain-gate junction leakages. By overlapping the gate into the substrate, the area of p-n (gate-channel) junction is reduced. Source-drain leakage is also reduced by reducing surface leakage around the gate. The threshold voltages of these devices was calculated by the extrapolation of the I_D vs

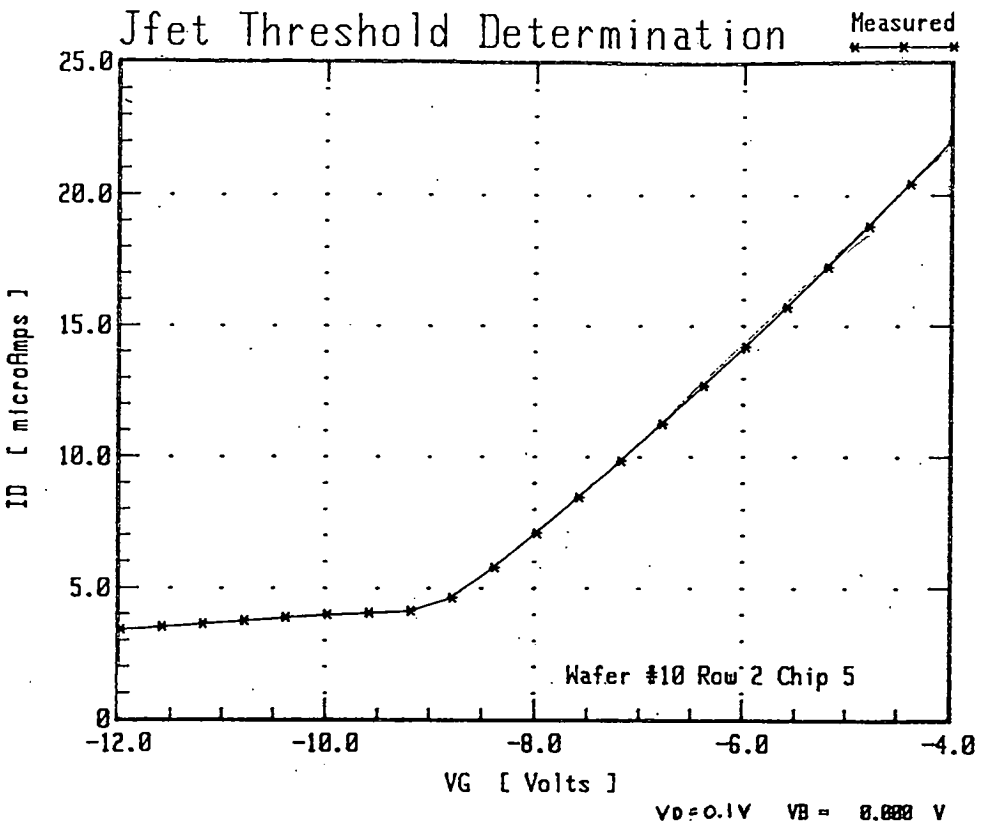


Figure 5.9a. Typical V_T extraction from non-overlap (device 1) JFET. This wafer had a 1 hr n-well drive in.

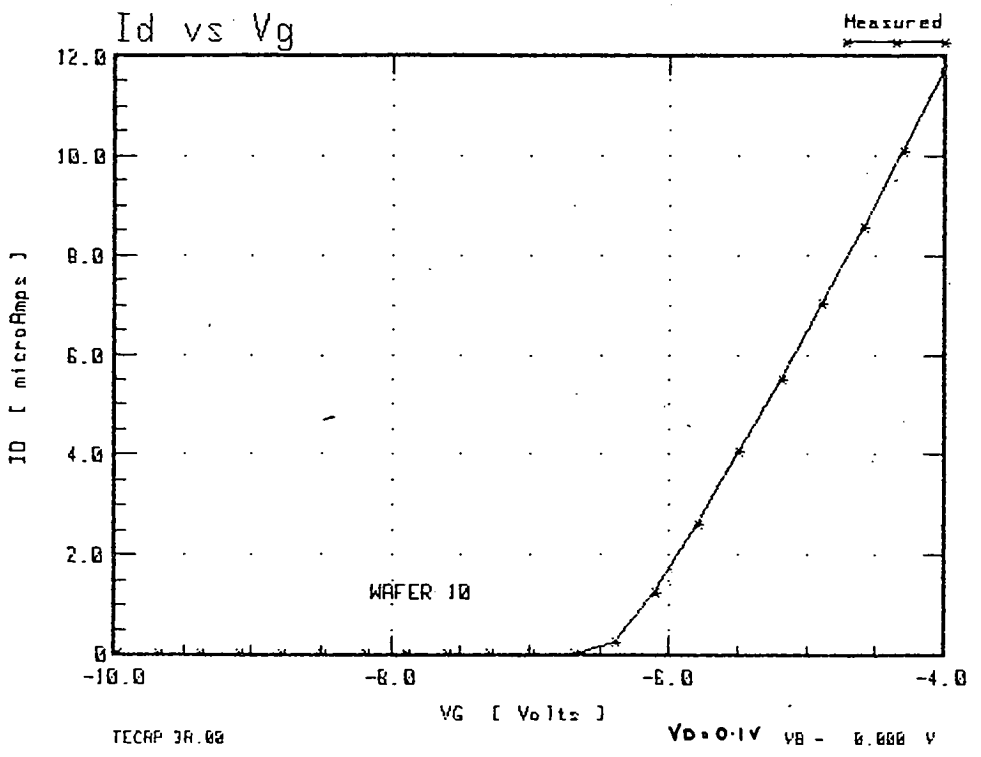


Figure 5.9b Typical V_T extraction from overlap (device 2) JFET. This wafer had a 1 hr n-well drive in.

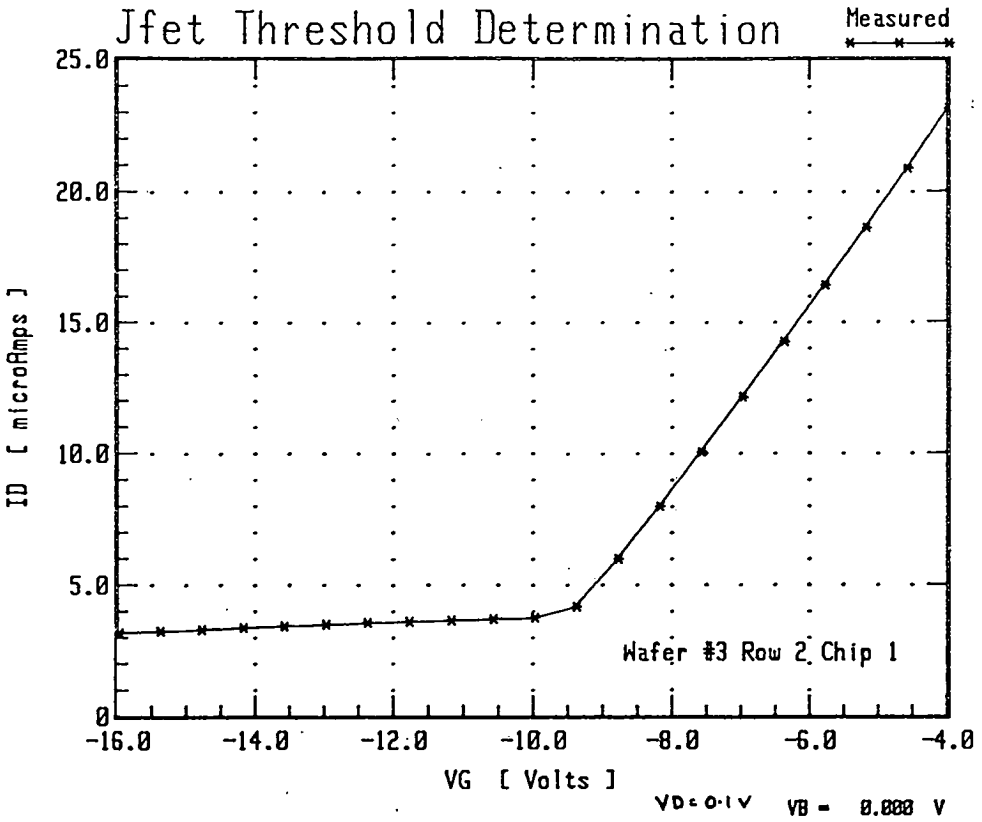


Figure 5.9c. Typical V_T extraction from non-overlap (device 1) JFET. This wafer had a 2 hr n-well drive in.

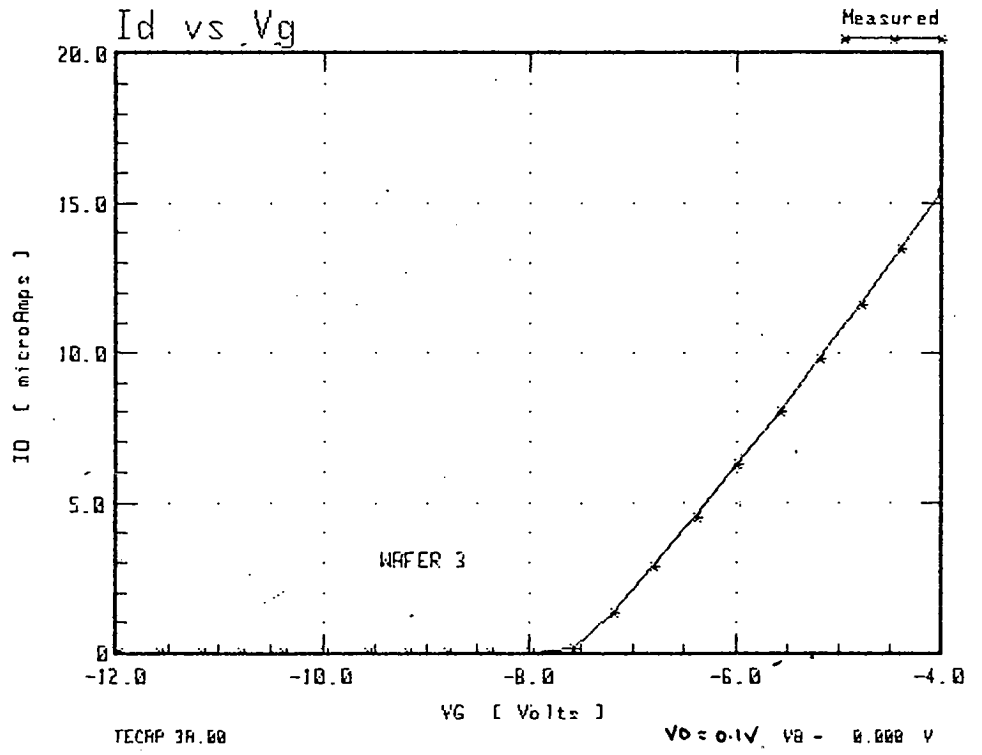


Figure 5.9d. Typical V_T extraction from overlap (device 2) JFET. This wafer had a 2 hr n-well drive in.

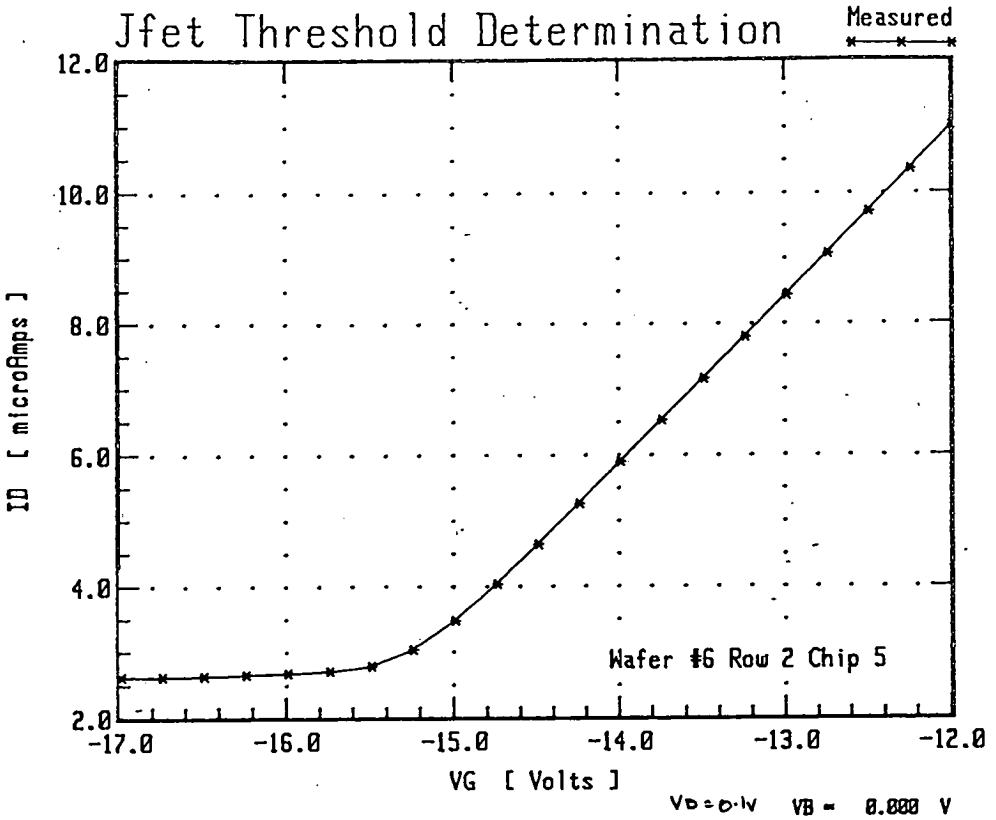


Figure 5.9e. Typical V_T extraction from non-overlap (device 1) JFET. This wafer had a 4 hr n-well drive in.

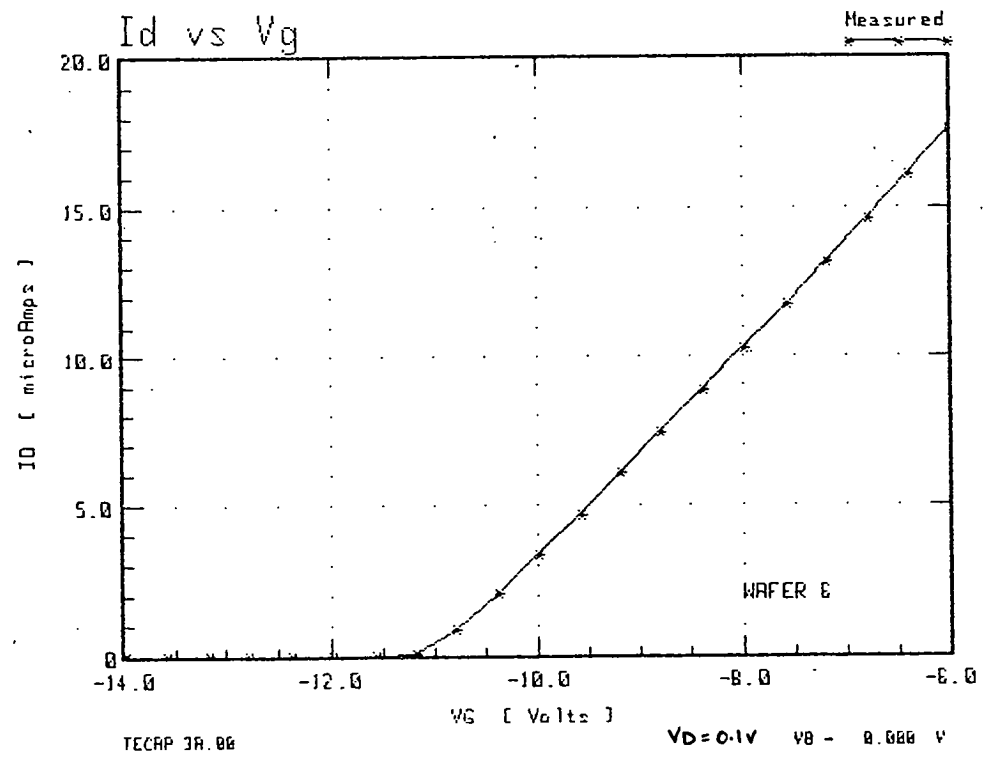


Figure 5.9f. Typical V_T extraction from overlap (device 2) JFET. This wafer had a 4 hr n-well drive in.

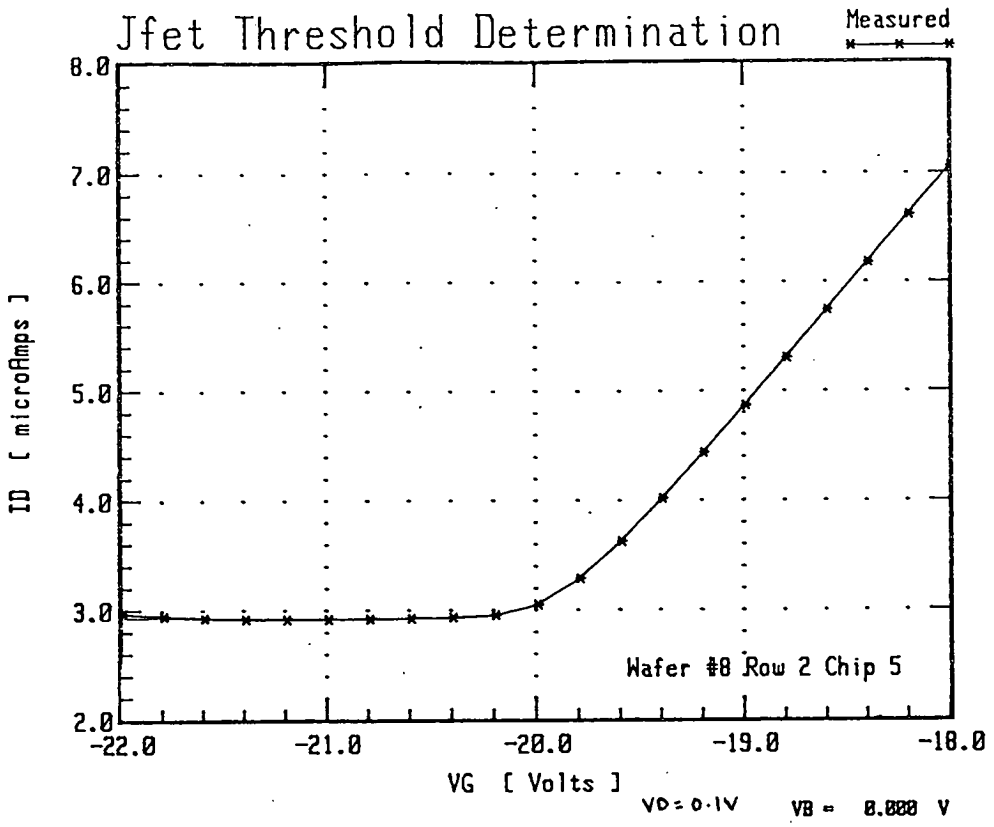


Figure 5.9g. Typical V_T extraction from non-overlap (device 1) JFET. This wafer had a 8 hr n-well drive in.

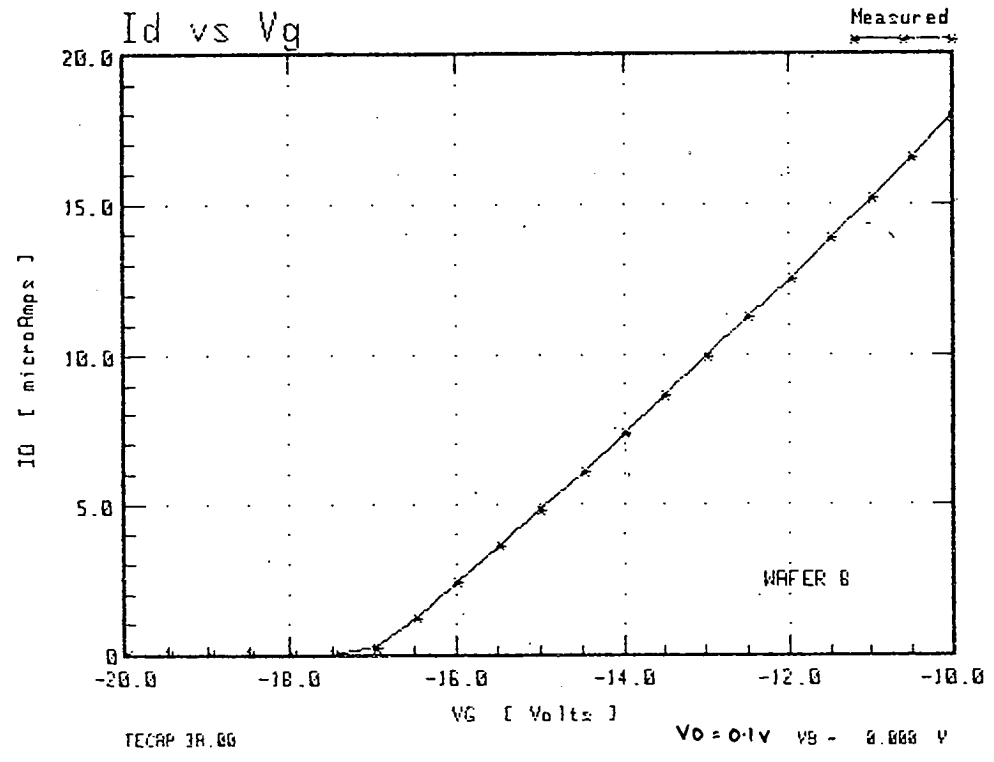


Figure 5.9h. Typical V_T extraction from overlap (device 2) JFET. This wafer had a 8 hr n-well drive in.

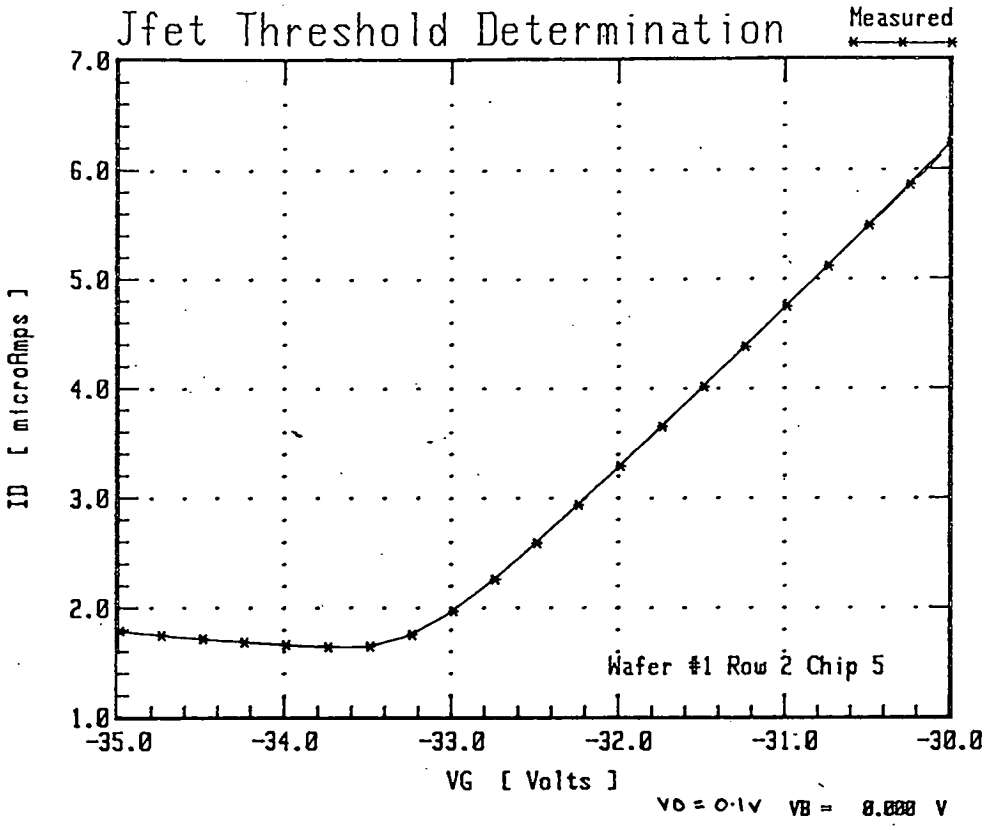


Figure 5.9i. Typical V_T extraction from non-overlap (device 1) JFET. This wafer had a 18 hr n-well drive in.

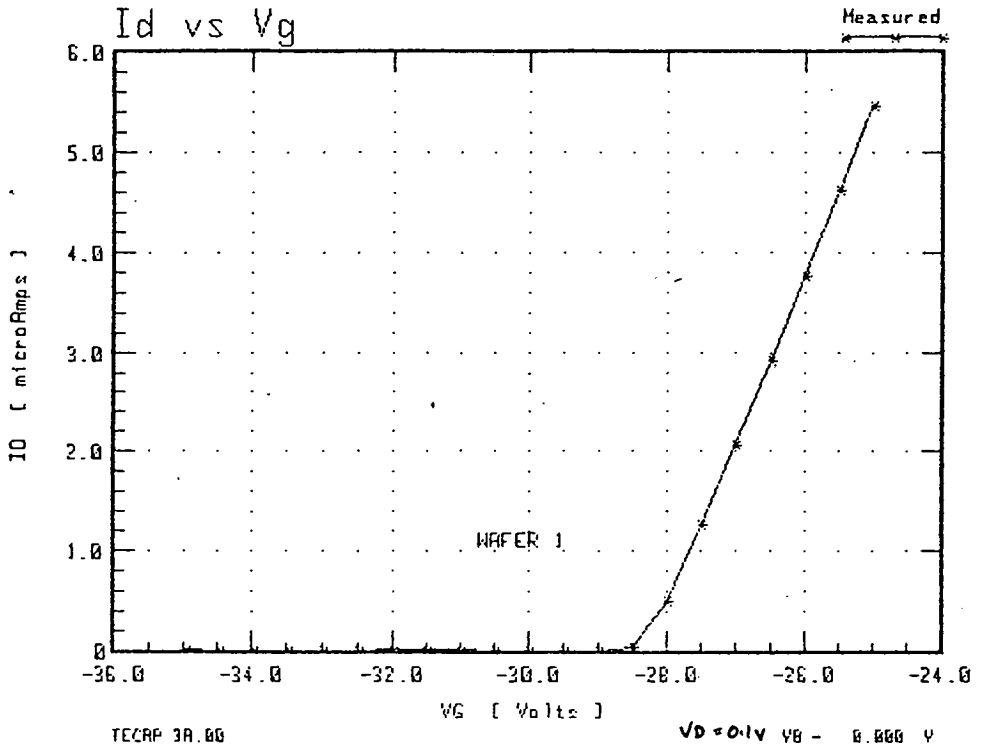


Figure 5.9j. Typical V_T extraction from overlap (device 2) JFET. This wafer had a 18 hr n-well drive in.

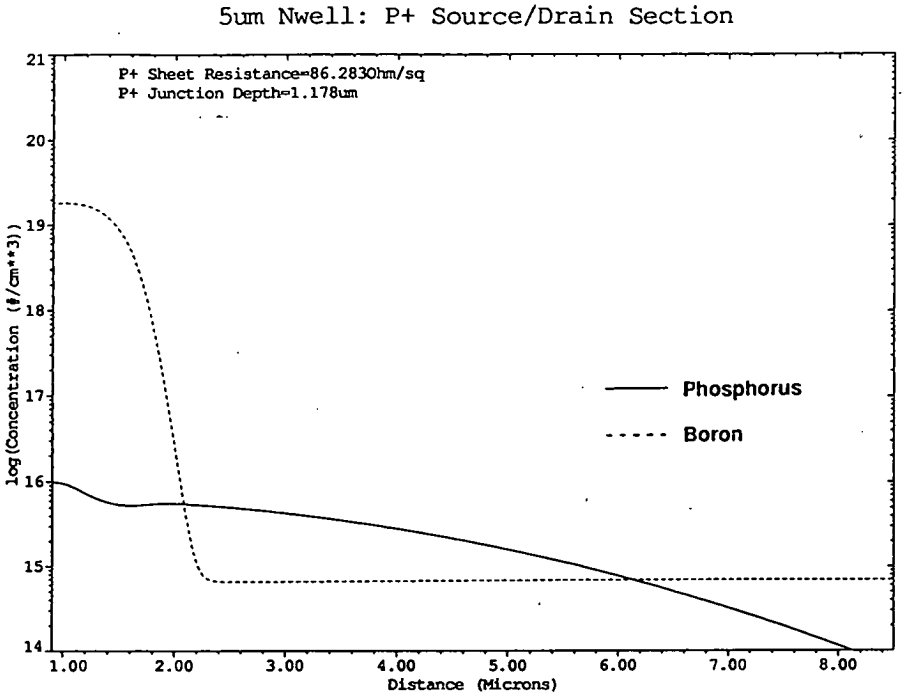


Figure 5.10 Typical SUPREM simulation of EMF 5μm process n-well diffusion (not taken from a process split).

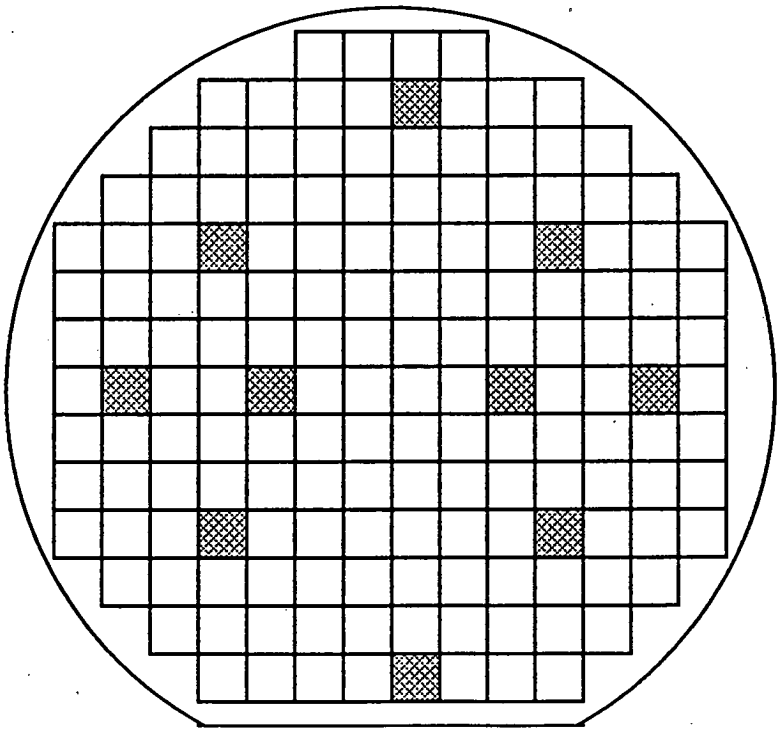


Figure 5.11 Wafer positions for 10 point average of extracted JFET V_T.

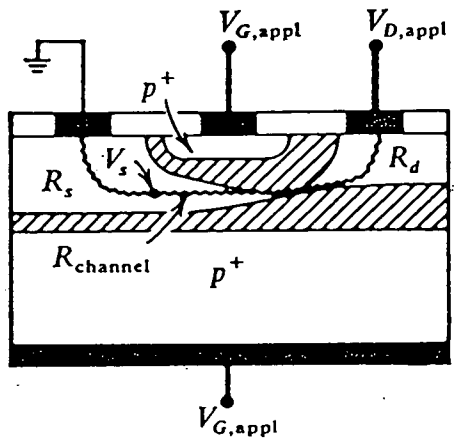


Figure 5.12 Channel region of a JFET with gate length L showing depletion regions when the substrate is clamped to the gate potential.

SUPREM simulated well depth	N-Well Diffusion time	Average V_T overlap (device 2)	Average V_T non-overlap (device 1)
2.92 µm	1	-6.4 V	-8.9 V
3.31 µm	2	-7.5 V	-9.5 V
3.92 µm	4	-11 V	-15.2 V
4.79 µm	8	-17 V	-19.9 V
6.04 µm	18	-28.3 V	-33.1 V

Table 5.1 Average threshold voltages extracted from overlapped vs non overlapped gates for different n-well drive in times.

V_G to zero drain current even though the channel is not completely turned off.

5.5 Relationship of n-well depth to JFET threshold.

An estimate of the n-well depth for each of these devices was obtained using SUPREM simulation. Figure 5.10 shows a typical SUPREM output for the n-well simulation.

An average V_T was obtained for each wafer by extracting V_T from 10 devices on the wafer. Figure 5.11 shows the locations of the devices used. The average threshold of the overlapped gate transistors was less than that of the enclosed gate (for any particular well depth) as shown in Table 5.1. This is due to the substrate clamping as discussed above. The depletion region associated with the channel-substrate junction encroaches more on the channel since the substrate behaves like a second (only less effective) gate.

Figure 5.13 shows a plot of extracted V_T vs simulated well depth for device type 1 with an $x=y^2$ relationship superimposed. It can be seen from this plot that the extracted V_T relationship with simulated well depth is modeled very well by V_T = (n-well depth)^2. Figure 5.14 shows the linear relationship between the square root of the extracted V_T value and the simulated well depth. Figure 5.15 shows the relationship between extracted V_T and the n-well diffusion time associated with the split. Once more a linear relationship is seen.

To explain the linear relationship observed between the extracted V_T and diffusion time we must examine the diffusion model for impurities in silicon. Using a simplified one-dimensional diffusion model which assumes constant diffusivity and constant total dopant (i.e ion implantation) the solution of Fick's second law for the initial condition

$$C(x,0) = 0 \tag{5.18}$$

and the boundary conditions

$$\int C(x,t)\,dx = S \tag{5.19}$$

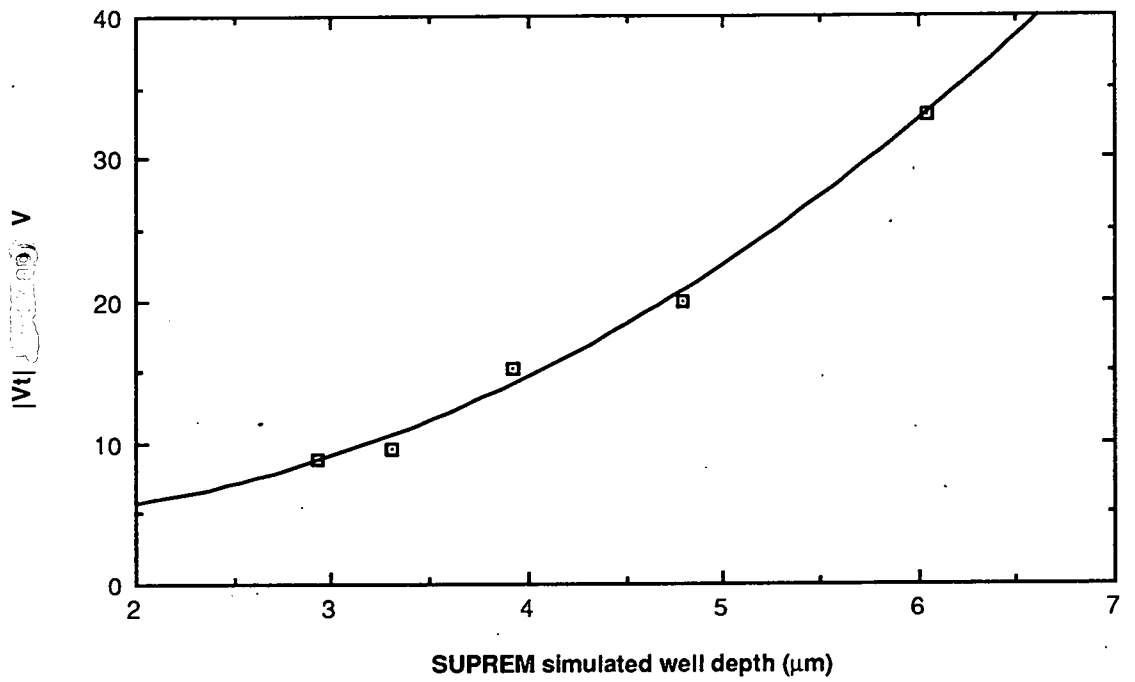


Figure 5.13 Measured V_T vs simulated well depth for JFET with non-overlapped gate (device 1), an $x=y^2$ relationship is superimposed.

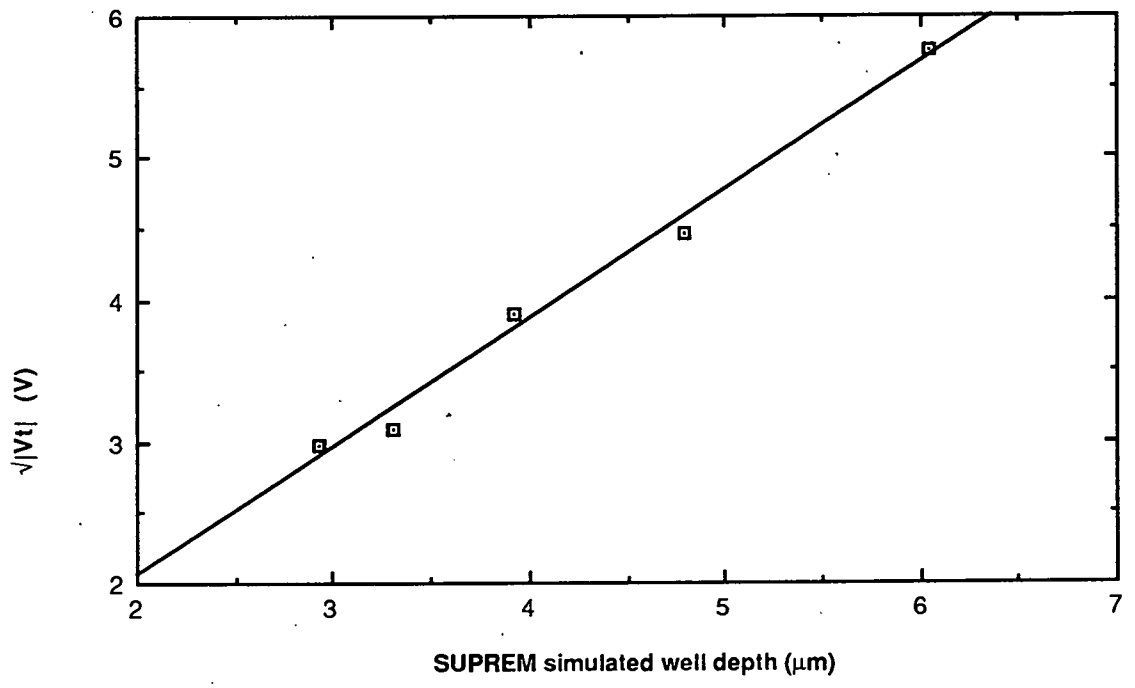


Figure 5.14 $\sqrt{V_T}$ vs simulated well depth for JFET (device 1).

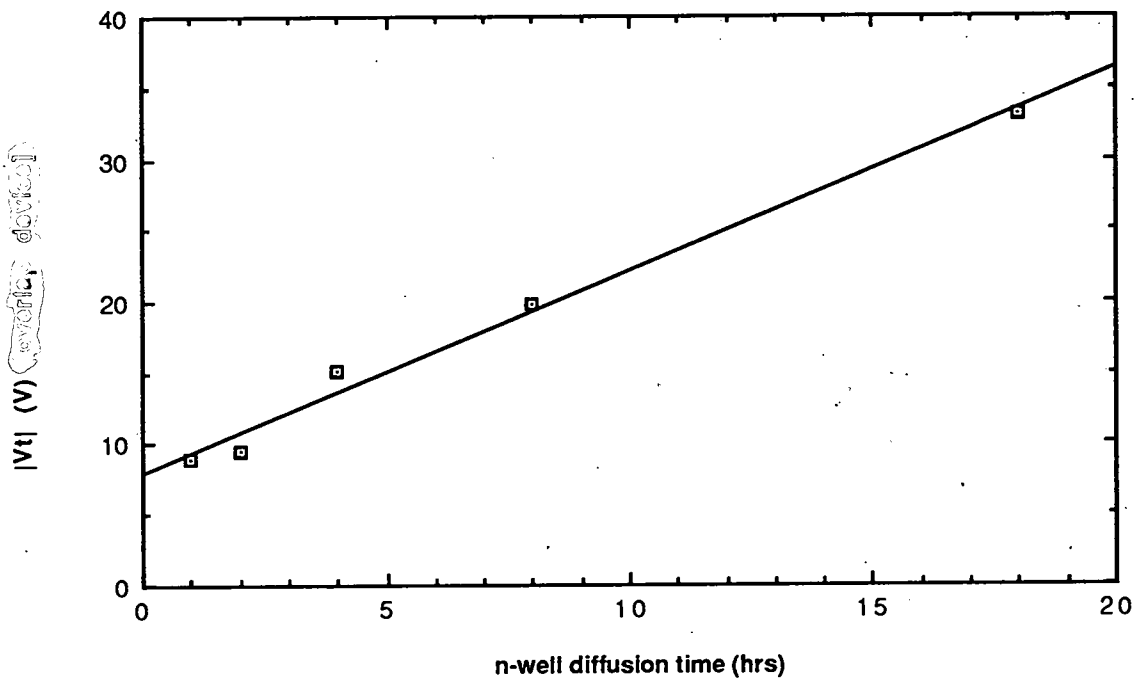


Figure 5.15 Measured V_T vs n-well diffusion time. (device type 1).

$$C(x,\infty) = 0 \tag{5.20}$$

is

$$C(x,t) = \frac{S}{\sqrt{\pi Dt}} exp\left[\frac{x^2}{4Dt}\right] \tag{5.21}$$

where C is the concentration of dopant (in atoms/cm^3) as a function of x, the depth into the silicon (in cm), D is the constant diffusion coefficient (in cm^2/s) and t is the diffusion time (in s). Using this model we can explain the linear dependence of V_T, the JFET threshold, with n-well diffusion time. The position where the diffusant concentration equals the substrate concentration is defined as the metallurgical junction x_j, that is $C(x_j,t) = C_{sub}$.

Thus

$$C_{sub} = \frac{S}{\sqrt{\pi Dt}} exp\left[\frac{x_j^2}{4Dt}\right] \tag{5.22}$$

and

$$x_j^2 = 4Dt\, ln\left(\frac{C_{sub}\sqrt{\pi Dt}}{S}\right). \tag{5.23}$$

We know from equation 5.12 that $V_T \propto t^2$ where t is the channel depth. If we approximate $t \approx x_j$, then $V_T \propto (t_{diff})$ where t_{diff} is the diffusion time associated with n-well drive in. This is only approximately true. We have assumed that the logarithmic term in equation 5.23 is dominated by the factor C_{sub}/S. and is constant for each of the n-well drive in times. However, this linear relationship is borne out by the extracted V_T (figure 5.15).

5.6 Cross wafer variation.

By using the mapping facility of TECAP it was possible to obtain

wafer maps of V_T. Figures 5.16a-e show wafer maps for each part of the n-well diffusion time split. Cross wafer patterns are readily observable for several wafers in Figure 5.16. This would suggest that a parameter such as V_T is indeed sensitive to CMOS process parameters. From the discussions above we can say that these variations may be due to effective n-well depth variation across the wafer. This non-uniformity may be caused by n-well dose variation or uneven thermal distribution across the wafer at the drive in stage. The non-uniformity may even be the result of incoming silicon doping variation.

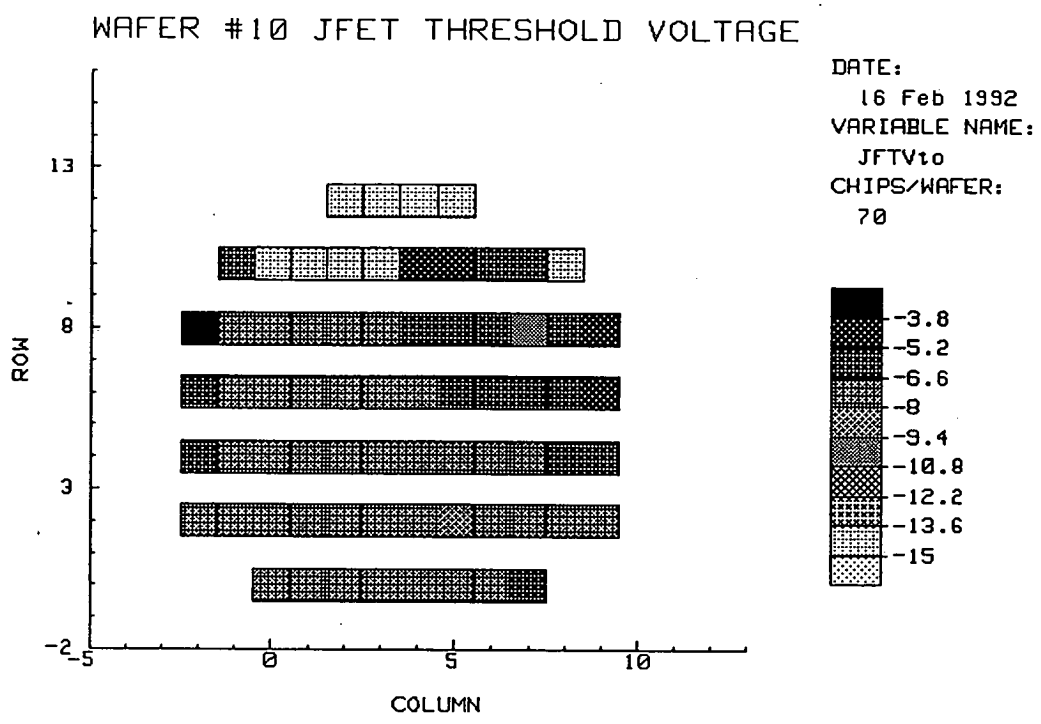
Chapter 3 discussed alternative methods of junction depth monitoring. These are almost all destructive techniques, bevel and stain etc, or relatively complex Capacitance-Voltage methods. This technique would enable well-junction monitoring in a standard CMOS process using simple and fast parametric extraction. When used in conjunction with other extracted parameters, a more complete understanding of process variation can be obtained.The relationship between JFET V_T and parameters extracted from parasitic bipolar transistors and MOS devices is discussed in chapter 7.

5.7 Conclusions.

JFET transistors are simply designed and easily fabricated in a standard CMOS process. Device characterisation is obtained from readily available commercial software. In-house parametric extraction routines can easily be modified to extract a few simple parameters from the JFET device, V_T, I_S. Because the structures require no extra masks they can be easily incorporated into existing process control chips or scribe grid process control devices. They reveal information about well depth and doping concentration. Diode information is inherent in the leakage currents associated with the device.

The parasitic JFET represents a device which could be easily utilised to enhance parametric information available for CMOS process control.

(a)



(b)

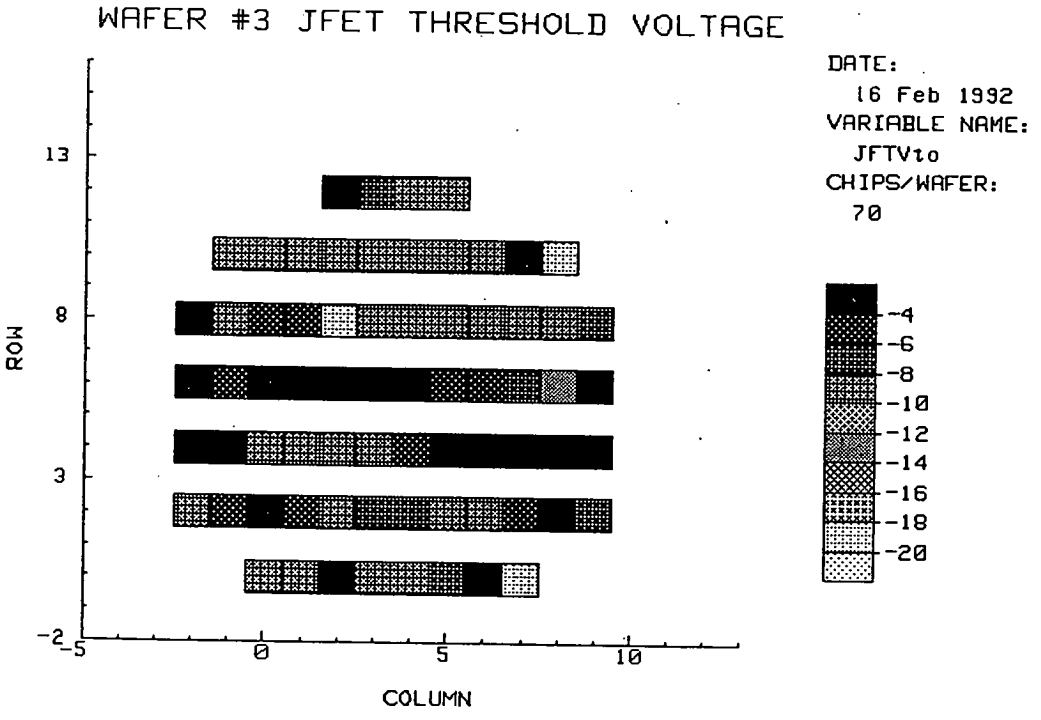
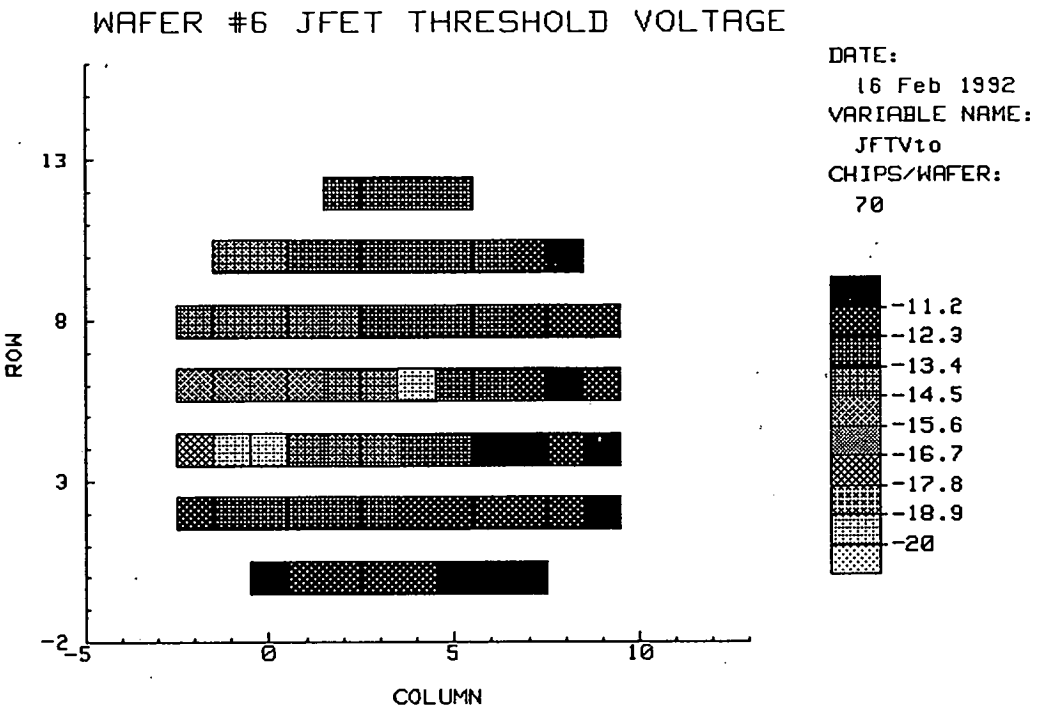


Figure 5.16 a-b Wafer maps of extracted V_T for (a) 1hr n-well drive in. (b) 2hr n-well drive in.

(c)



(d)

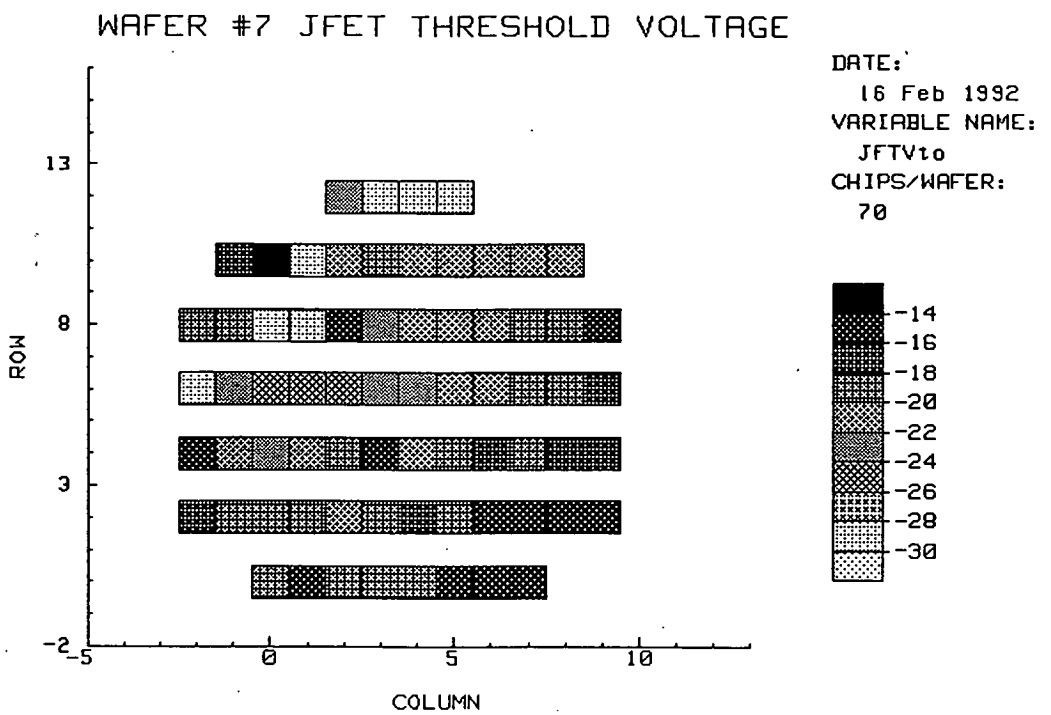


Figure 5.16 c-d Wafer maps of extracted V_T for (c) 4hr n-well drive in. (d) 8hr n-well drive in.

(e)

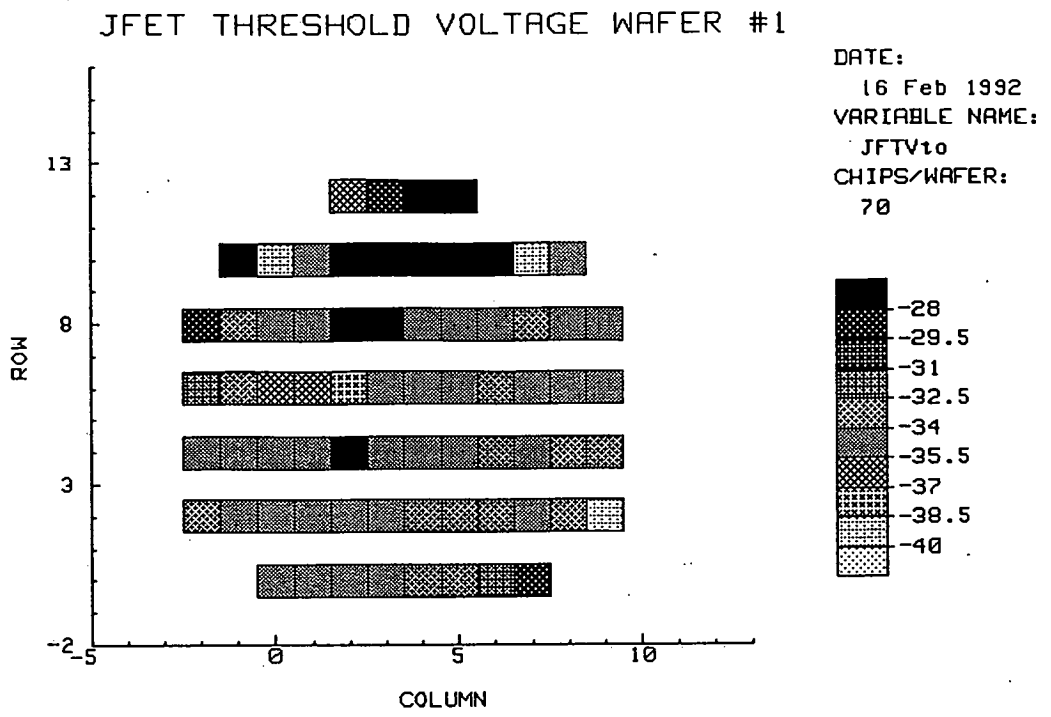


Figure 5.16 e Wafer map of extracted V_T for (e) 18hr (std) n-well drive in.

References

[1] S.M. Sze,"*Physics of Semiconductor Devices,*" 2nd Edition Wiley-Interscience, New York, p. 312, 1981.

[2] A.S. Grove,"*Physics and Technology of Semiconductor Devices,*" Wiley, p. 243, 1967.

[3] R.M Warner and B.L Grung,"*Transistors Fundamentals for the Integrated-Circuit Engineer,*"Wiley-Interscience, p. 754, 1983.

[4] R.S. Muller and T.I. Kamins, "*Device electronics for Integrated Circuits,*"Wiley 2nd. Ed., p. 204, 1986.

[5] S. M. Sze, "*Physics of Semiconductor Devices,*" 2nd Edition Wiley-Interscience, New York, pp323, 1981.

Chapter 6

ΔL Extraction Using Parasitic Bipolar Transistors

6.1 Introduction

Lateral bipolar transistors can be simply fabricated using a standard CMOS process. Using an n-well process with a boron doped source and drain and a phosphorus doped well, the lateral pnp bipolar device is easily fabricated. The performance of these bipolar devices is poor, since these structures have inefficient lateral transport of holes from emitter to collector which causes low common emitter gain and increases charge storage times [1]. Improvements can be made in the performance of lateral bipolar devices by the inclusion of an epitaxial layer and a heavily doped buried layer. However, this means lateral bipolar transistors can no longer be fabricated in a standard CMOS process. This chapter details a useful test structure for the extraction of the MOS sideways diffusion parameter using the simple lateral bipolar transistors. This is a measure of the amount of sideways diffusion of the source-drain regions under the gate. ΔL extraction is most commonly performed by arranging the Ianthola and Moll [2] equation for MOS current in the linear region

$$I_D = \beta V_G \left[V_G - V_T - \frac{1+F_b}{2} V_D \right] \tag{6.1}$$

where

$$\beta = \frac{(\varepsilon_{ox}\,\varepsilon_o)}{t_{ox}} \frac{W}{L} \cdot \mu_{eff} \cdot \tag{6.2}$$

L, the actual channel length can be described as

$$L = L_m - 2\Delta L \tag{6.3}$$

where L_m is the drawn length and ΔL the sideways diffusion. The expression for transistor gain β can be manipulated to obtain

$$\frac{1}{\beta} = \frac{L_m}{\mu_{eff} C_{ox} W_m} - \frac{2\Delta L}{\mu_{eff} C_{ox} W_m}. \tag{6.4}$$

If $1/\beta$ is then plotted against L_m, the intercept of the best fit straight line with the L_m axis is $2\Delta L$. It is important that the drain voltage is kept low to minimise the depletion layer around the drain. A large depletion layer would increase the measured value of ΔL. The gate voltage must be high enough to bias the transistor in the linear region, in order to maximise β and reduce the influence of parasitic source and drain contact resistances. The accuracy of this method is largely determined by the measurement made on the smallest drawn length transistor. Short channel effects such as the modulation of V_T give rise to increased values of β being calculated and a consequent reduction in the accuracy of the measurement.

The proposed use of lateral bipolar transistors avoids these inherent errors.

6.2 Theory of Lateral Transistor Operation

Lindmeyer and Schneider [3] analysed the simple model shown in Figure 6.1 providing expressions for both the lateral and vertical contribution of I_c. The collector current I_{cl} can be represented by

$$I_{cl} = p'_{ne} \frac{D_p}{W_{bo}} l_e x_j \tag{6.5}$$

where

$$p'_{ne} = \frac{n_i^2}{N_d}\left[exp\left(\frac{qV_{EB}}{kT}\right)-1\right]. \tag{6.6}$$

p'_{ne} is the excess hole concentration, D_p is the low level hole diffusion coefficient, W_{bo} is the drawn base width, l_e is the emitter length and x_j is the junction depth.

The simple bar structure must obviously be modified to represent a lateral pnp device fabricated in a standard CMOS process, this is shown in Figure 6.2. If the effects of curved diffused junctions are included then the collector current can be expressed as

$$I_c = I_{cl} F_G\left(\frac{W_{bo}}{x_j}\right) \tag{6.7}$$

where F_G is a geometrical factor representing the effect of structure geometry on I_C.

The geometrical factor for a lateral transistor fabricated using epitaxy and n+ buried layers was first solved by Chou [4] using 2D computer analysis. This function is shown in Figure 6.3. An analytically calculated value of the geometrical factor was presented by Seo and Kim [5]. Their method is used here to develop F_G. Figure 6.4 shows the geometry of the lateral p+ n p+ transistor to be considered. Three approximations are made :

1. The collector current flows horizontally when W_{bo} is not much greater than x_j.

2. The edge profile of the emitter and collector diffusions are circular.

3. The doping concentration of the n-well from $x = 0$ until $x = x_j$ is constant.

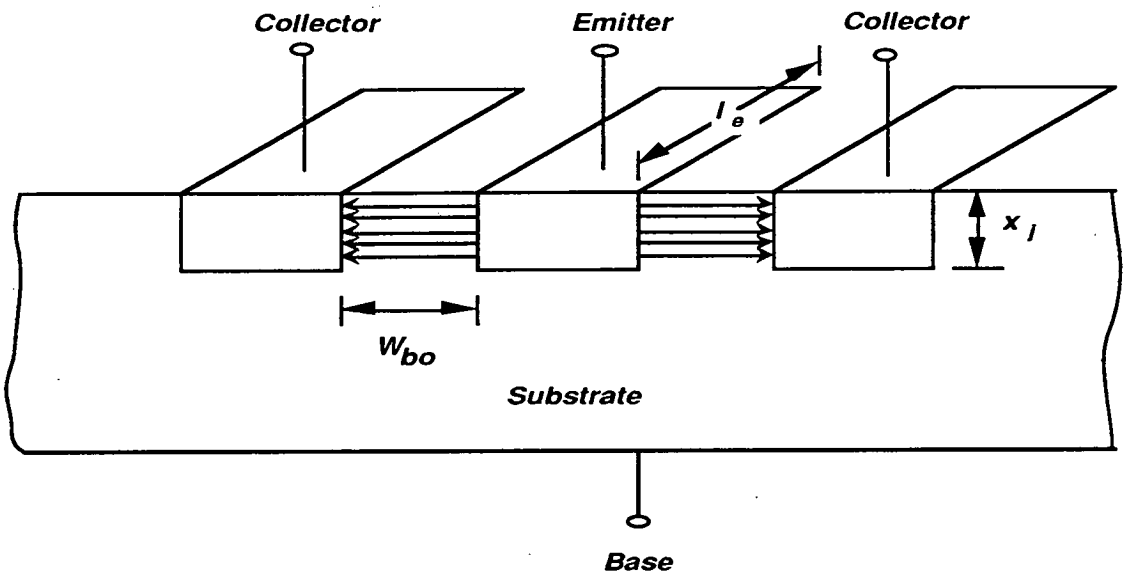


Figure 6.1. Simple model of lateral transistor used to obtain expressions for lateral and vertical currents.

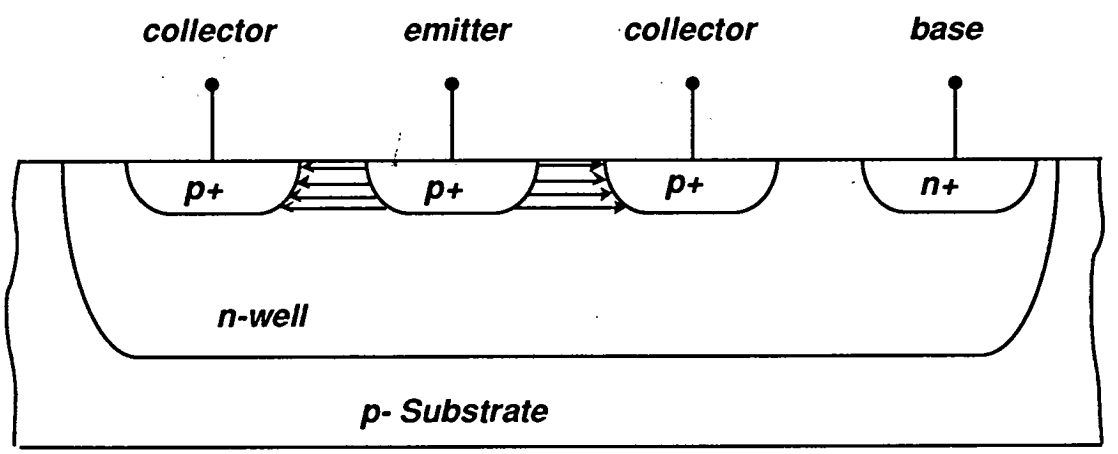


Figure 6.2 Schematic of simple lateral bipolar transistor fabricated using a n-well CMOS process.

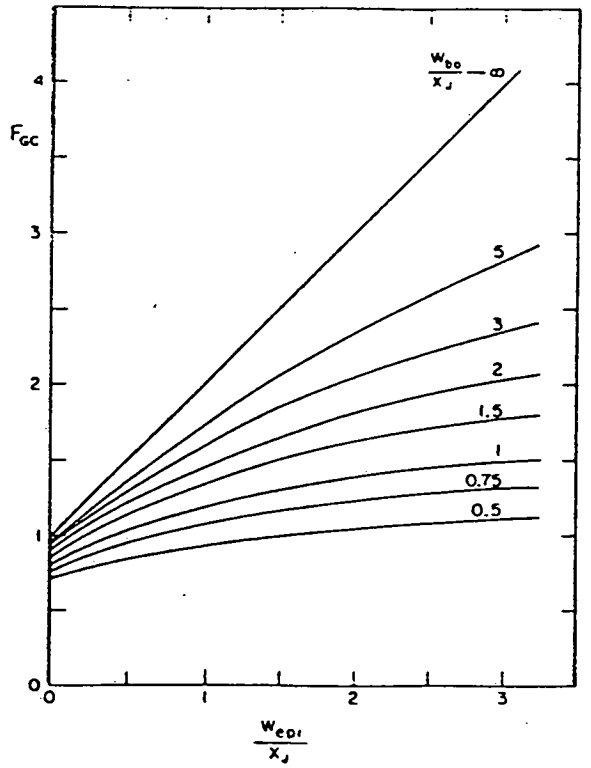


Figure 6.3 The geometrical factor as a function of x_j.

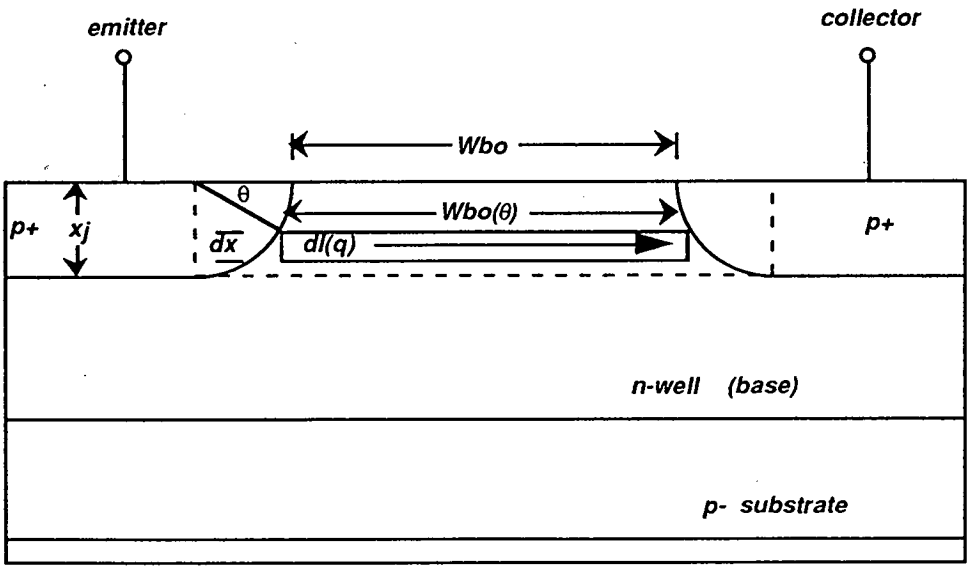


Figure 6.4 Schematic of the lateral p⁺-n-p⁺ transistor to be analysed.

In Figure 6.4 the emitter length is l_e. The base width at angle θ is represented as

$$W_b(\theta) = (W_{bo} + 2x_j) - 2x_j\cos(\theta) \quad . \tag{6.8}$$

The current $dI_1(\theta)$ now flowing in a stripe whose width is $dx(\theta)$ is given by

$$dI_1(\theta) = qp'_{ne}\frac{D_p}{W_b(\theta)}\, xl_e\, dx(\theta) = K\frac{dx(\theta)}{W_b(\theta)} \tag{6.9}$$

where

$$d\,x(\theta) = x_j\cos\theta d\,\theta \tag{6.10}$$

and

$$K = qp'_{ne}D_p l_e \quad . \tag{6.11}$$

From equations 6.9 and 6.10 the total current I_c can be expressed as

$$I_c = \int_0^{\frac{\pi}{2}} dI_1(\theta) \tag{6.12}$$

$$I_c = K\left[\frac{\pi}{4} + (W_{bo} + 2x_j)\frac{1}{\sqrt{W_{bo}^2 + 4W_{bo}x_j}}\right.$$

$$\left. tan^{-1}\frac{W_{bo} + 4x_j}{\sqrt{W_{bo}^2 + 4W_{bo}x_j}}\right]. \tag{6.13}$$

Then the analytic geometry factor

$$F_G = \frac{I_c}{I_{cl}} \tag{6.14}$$

$$= y \left[-\frac{\pi}{4} + (y+2) \frac{1}{\sqrt{y^2+4y}} tan^{-1} \frac{y+4}{\sqrt{y^2+4y}} \right] \quad (6.15)$$

where

$$y = \frac{W_{bo}}{x_j}. \quad (6.16)$$

6.3 ΔL Extraction

If ΔL, the sideways diffusion associated with MOS transistors, is incorporated into the geometry of the lateral pnp bipolar, then the transistor's geometry can be represented by Figure 6.5. The expression 6.13 developed for I_{cl} becomes

$$I_{cl} = K \left[-\frac{\pi}{4} + ((W_{bo}-2\Delta L)+2x_j) \frac{1}{\sqrt{(W_{bo}-2\Delta L)^2 + 4(W_{bo}-2\Delta L)x_j}} \cdot \right.$$

$$\left. tan^{-1} \frac{W_{bo}-2\Delta L+4x_j}{\sqrt{(W_{bo}-2\Delta L)^2 + 4(W_{bo}-2\Delta L)x_j}} \right]. \quad (6.17)$$

To extract ΔL, an array of transistors with a number of different drawn W_{bo} dimensions must be biased for low level injection with the same V_{EB}. This is the case when $V_{EB} \ll V_{HL}$ [4] and V_{HL} is given by

$$V_{HL} = \frac{2kT}{q} ln \frac{N_d}{n_i}. \quad (6.18)$$

Then if $| 1/I_{cl} |$ is plotted against drawn value of W_{bo} the intersection of the extrapolated line with the W_{bo} axis is $\approx 2\Delta L$.

6.4 Simulation

Using the expression for I_{cl} developed above (6.15), graphs of $1/I_{cl}$ versus drawn W_{bo} can be simulated for arrays of transistors biased identically. From (6.15) various values of ΔL can be introduced into the equation. Figures 6.6 and 6.7 show these simulated curves for two values

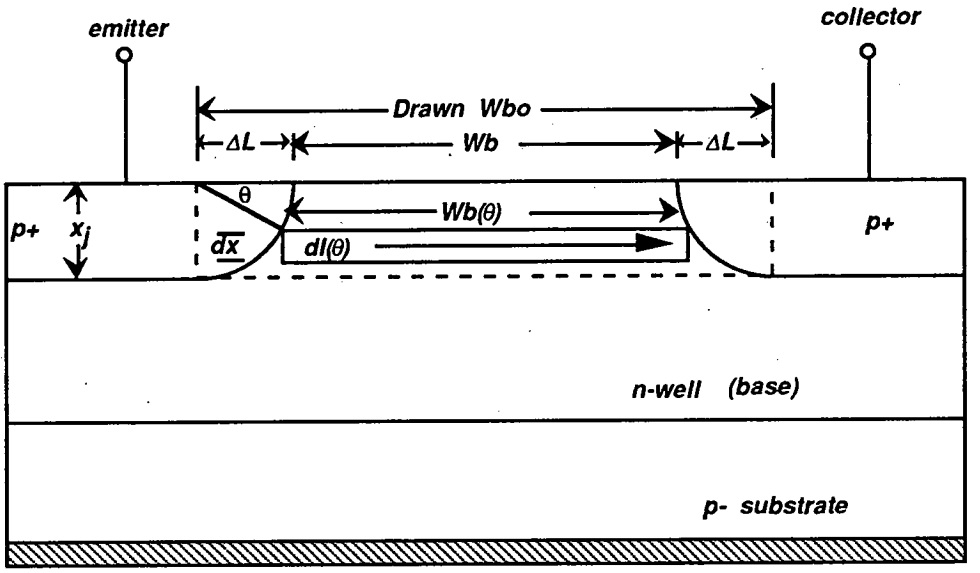


Figure 6.5 Schematic of the lateral p^+-n-p^+ transistor with the effect of ΔL the sideways diffusion included.

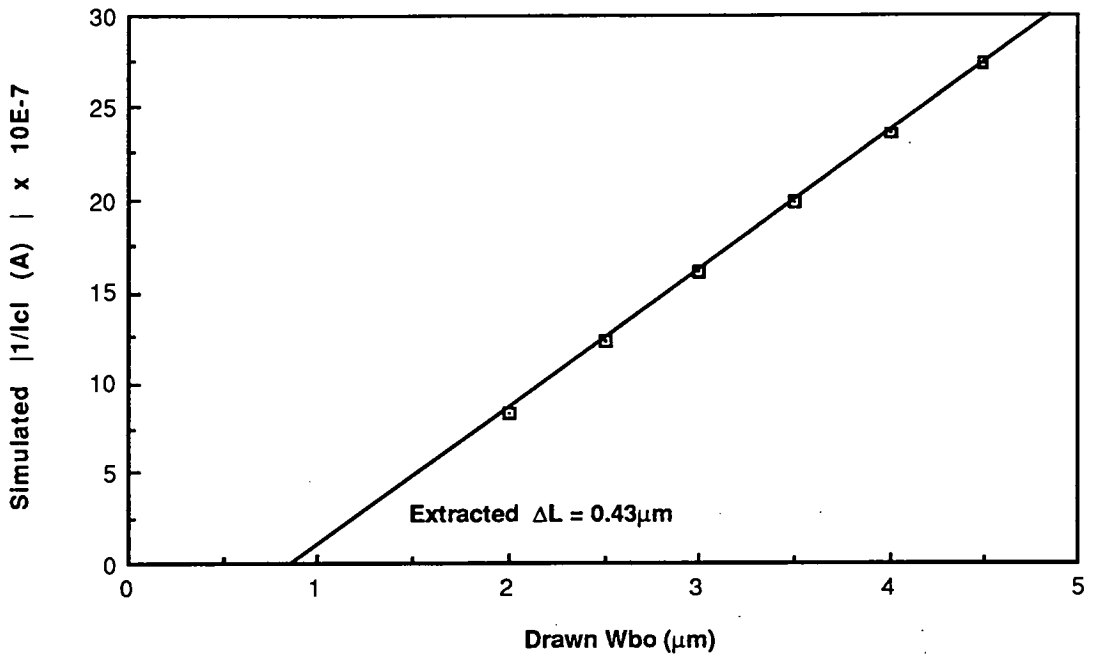


Figure 6.6 Simulated data for a ΔL of 0.5μm. Regression shows extrapolated value of ΔL.

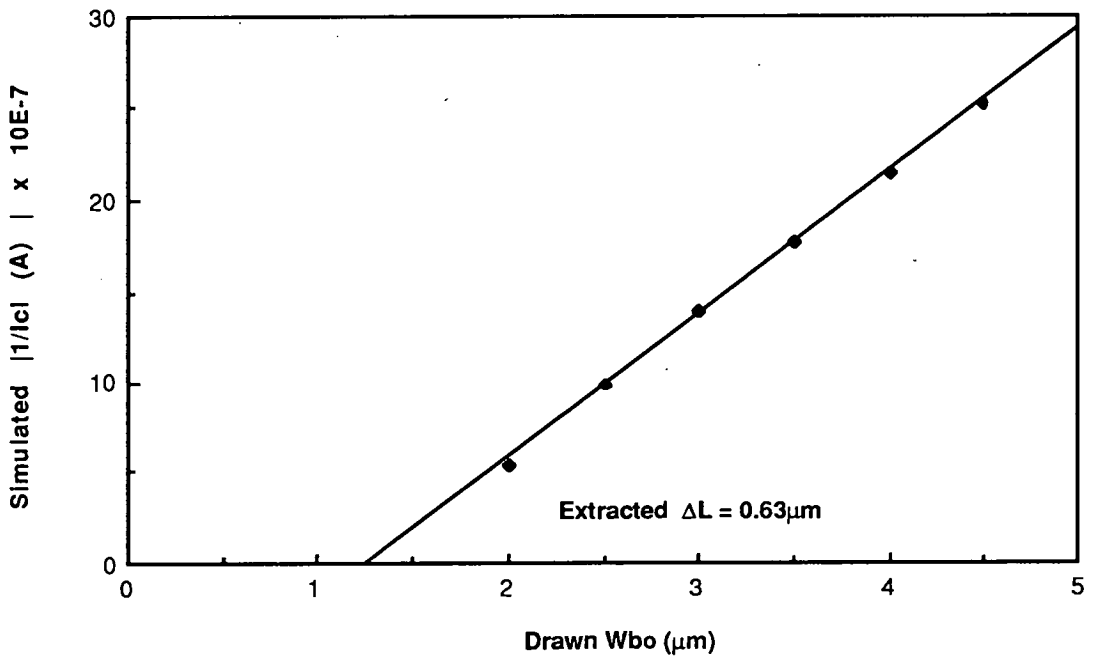


Figure 6.7 Simulated data for a ΔL of 0.7μm. Regression shows extrapolated value of ΔL.

of ΔL, 0.5μm and 0.7μm.

Using linear extrapolation, estimates of ΔL can be achieved from the five simulated values on the graph. This is for the worst case sideways diffusion in 6.17 where $\Delta L = x_j$. In the case of implanted source, drain transistors with shorter heat treatment, when $\Delta L < x_j$, the transistor structure becomes more like the simple bipolar in Figure 6.1. Thus the relationship between I_{cl} and W_{bo} becomes more linear and the accuracy of regression is increased.

6.5 Experimental Method

To verify the physical operation of this method, lateral bipolar transistors were incorporated into the design of the two test chips discussed in chapter 4. These fell into two groups; bar type emitter-collector structures and circular emitter-collector structures. Figures 6.8a-c show schematics of these. The circular transistors were included to reduce edge effects and provide a completely closed emitter. Problems arose, however, at the mask making stage and the circular transistors were flawed with several of the emitters being shorted to the collectors.

Thus the characterisation involved in this experiment was limited to the bar-type structures. Each of the devices shown in figures 6.8 a-b was fabricated with drawn W_{bo} values from 2μm to 4μm in 0.5μm steps. Identical npn transistors were fabricated and characterised.

6.5.1 Device Processing

The devices were processed using the standard EMF 5μm n-well CMOS process. The details of this process are given in chapter 4.

6.6 Electrical Characterisation

Figure 6.9 shows the arrangement of the SMUs used for the characterisation of the lateral bipolar devices designed for this experiment bipolar device. The detailed electrical characterisation of a typical lateral device is given below. This transistor was chosen only to illustrate the method used on many transistors on each wafer.

6.6.1 R_E and R_C Extraction

The measured data used to extract R_E and R_C is of the form collector current vs collector voltage at several base currents. Figure 6.10 shows a sweep measurement of collector current vs collector voltage, it indicates the two distinct regions of operation. Region one is when the device is in forward active operation, region two is when the device is in reverse saturation. $1/R_C$ is measured as the slope of these lines, as can be seen from figure 6.10. R_C is smaller than $R_{Cnormal}$ (normal is used here as most devices are operated in the forward active region). A value of R_E is extracted from these curves in the saturated region.

$$R_E = \frac{\Delta V_{CE}}{I_{b2} - I_{b1}} \tag{6.19}$$

As many measurements can lead to device heating and subsequent altering of device characteristics, a limited region must be chosen in which to measure R_E and R_C. This region can be the same for both cases and the measurements can be made simultaneously. This type of measurement, however, is only really applicable to devices with more heavily doped collector regions where the value of R_{Csat} and $R_{Cnormal}$ are almost the same, i.e. the value of R_C does not vary significantly with collector current.

The parasitic device characterised here, however, has a relatively lightly doped collector (p+ source drain doping) and two separate regions must be chosen for the extraction of R_E and R_C. Tables 6.1 and 6.2 indicate the regions chosen to extract R_E and R_C from the device under test (DUT). Fig 6.11 and 6.12 show the resulting measured curves. Values of 6.2 ohms and 241 ohms were extracted for R_E and R_C respectively.

6.6.2 V_{AF} and V_{AR} Extraction

V_{AF} and V_{AR} are parameters which characterise the effect of base width modulation on the transistor caused by variation of the collector-base space charge layer (the Early effect). The Early effect was discussed

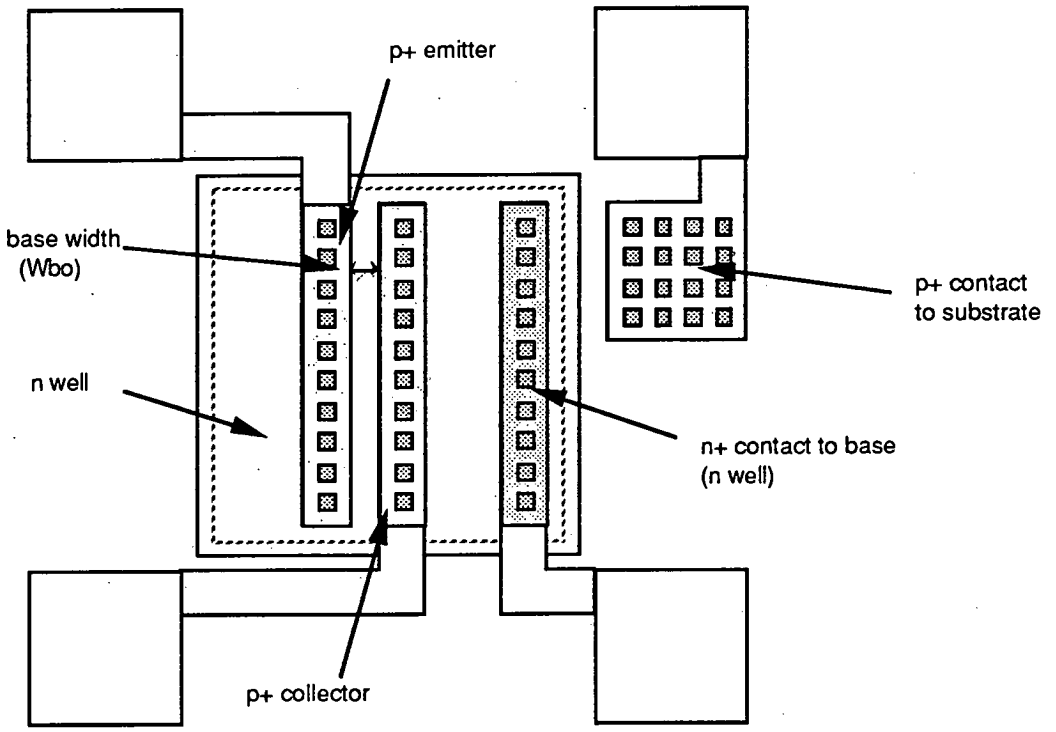


Figure 6.8a Single stripe emitter with single stripe collector. One of the simplest lateral designs.

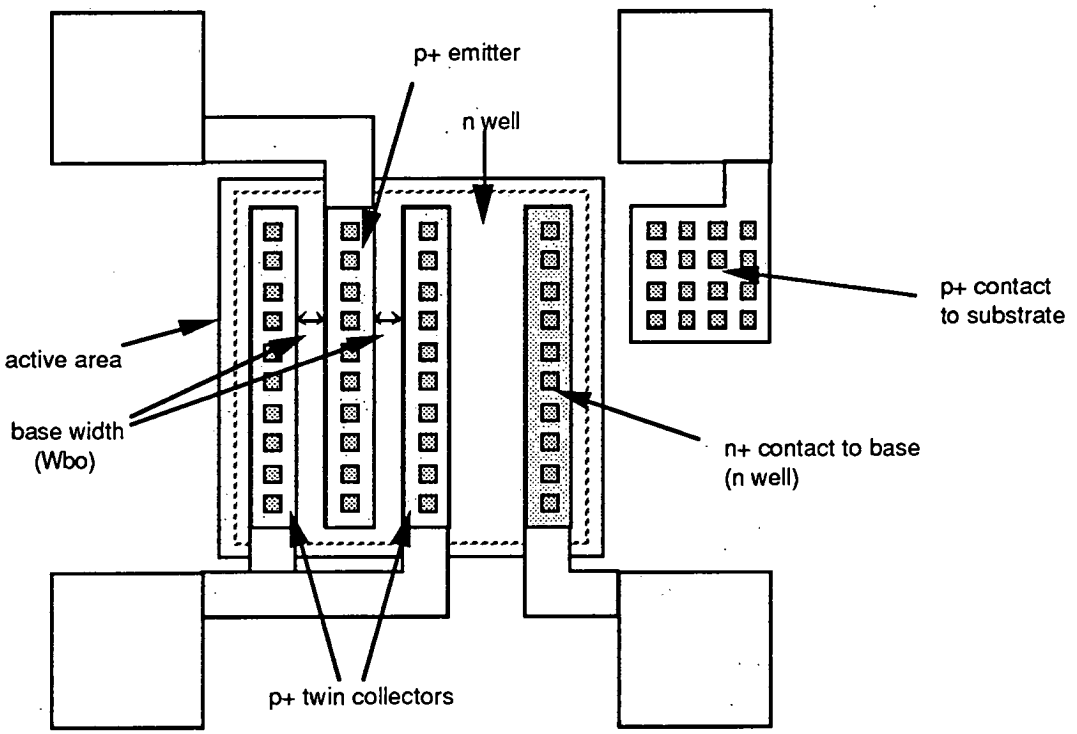


Figure 6.8b Single stripe emitter with twin stripe collector. This structure is most like the simple lateral transistor in the analysis.

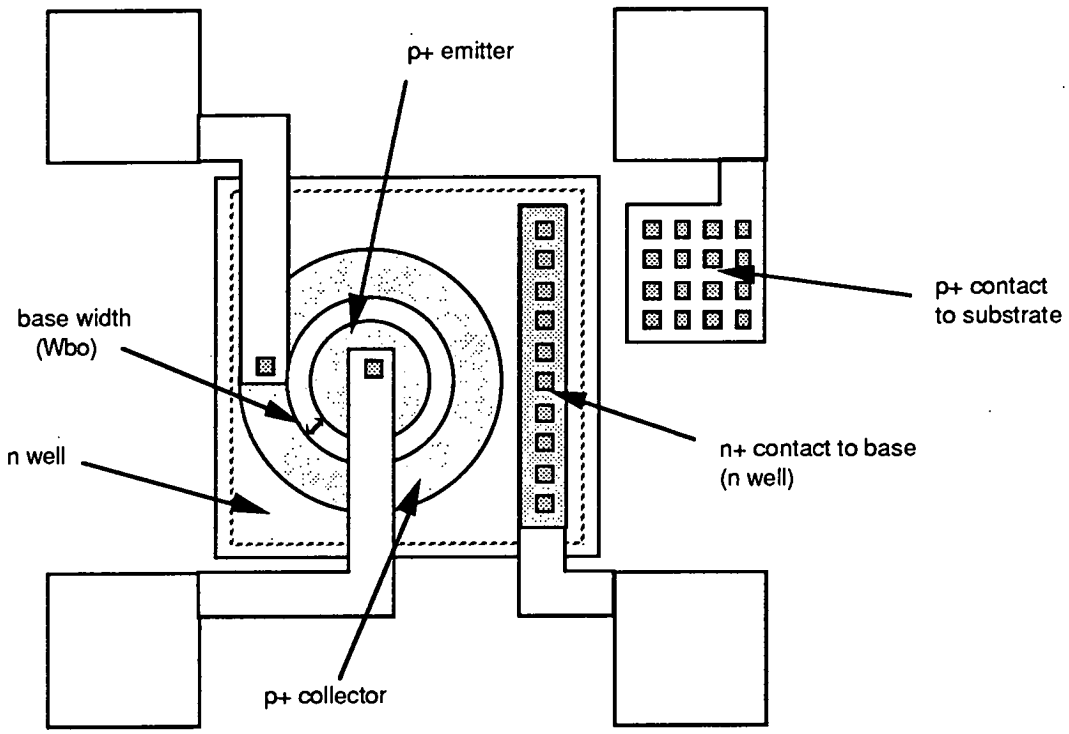


Figure 6.8c Circular transistor with totally enclosed emitter. This transistor was designed to avoid edge effects.

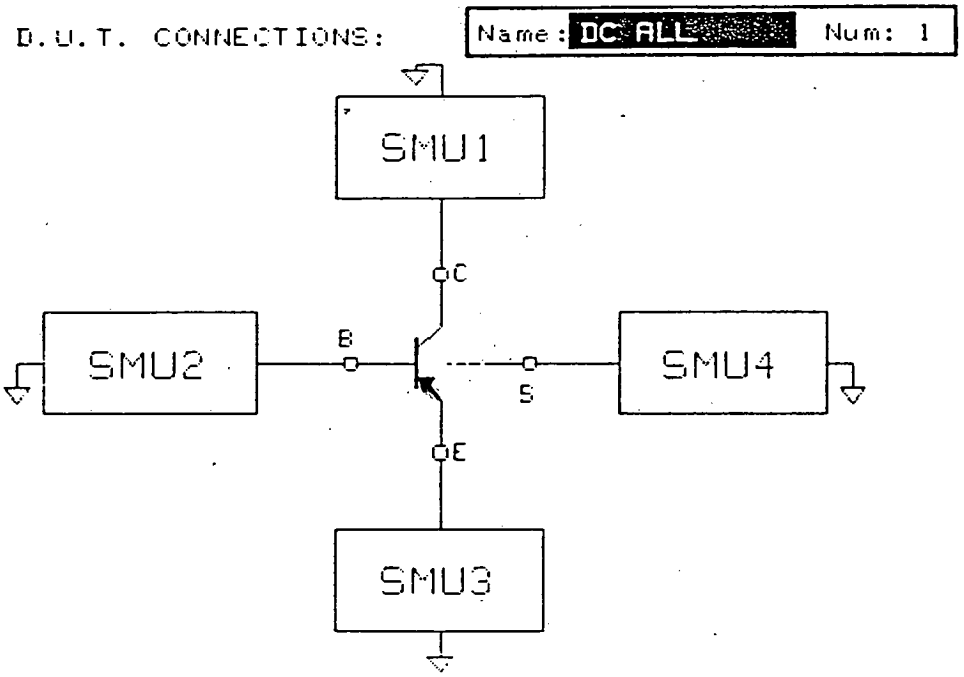


Figure 6.9 SMU arrangement for the characterisation of the lateral structures.

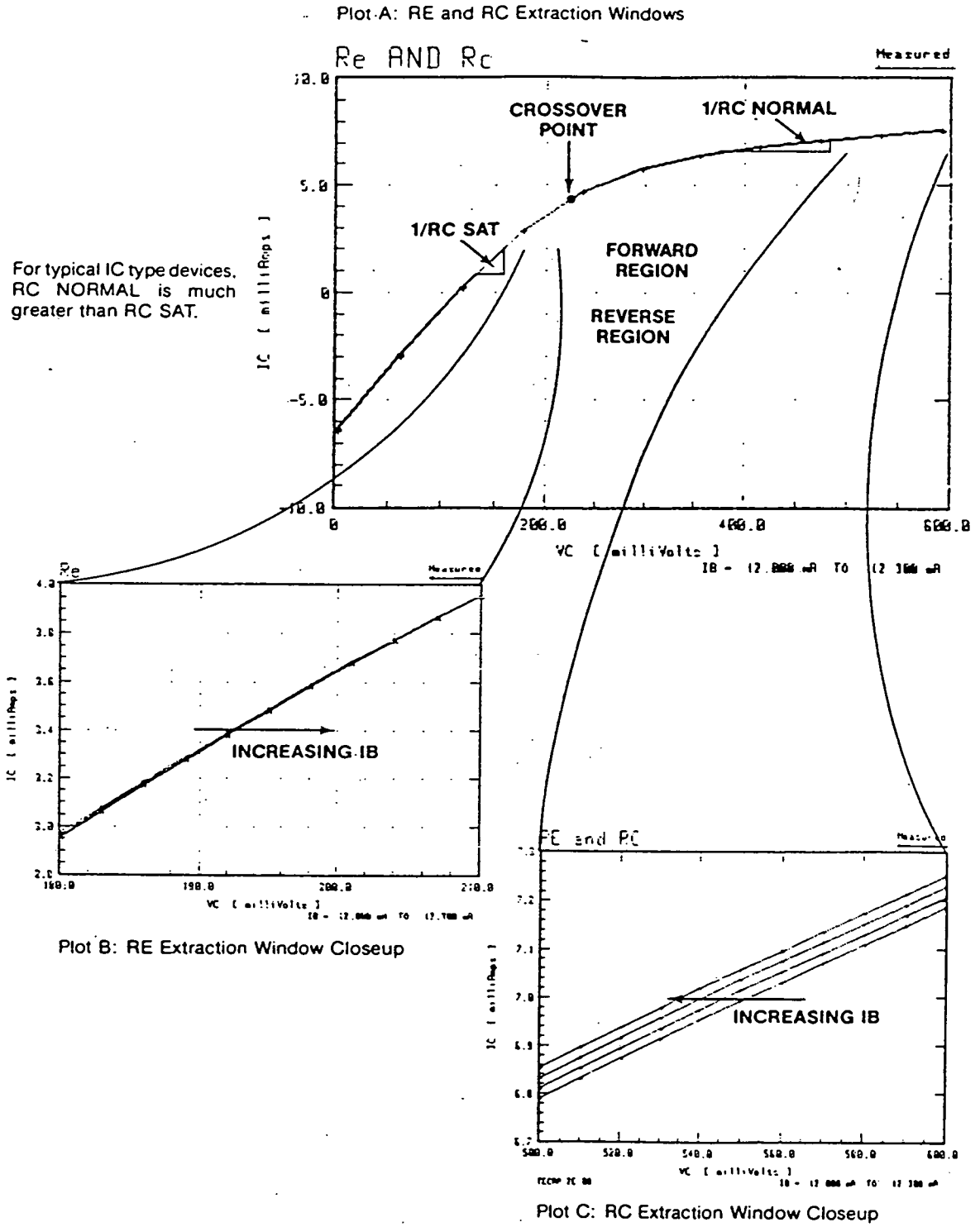


Figure 6.10 Sweep of collector current vs voltage showing the region of extraction for R_C and R_E.

Setup Name : Setup # 1 RE			
Function	MAIN	STEP	CONSTANT
Source Name	VC	IB	VE
Sweep Mode	LIN	LIN	
Start	-120 mV	-10 mA	0.0 V
Stop	-160 mV	-10.3 mA	
# of Points	11	4	
Compliance	100 mA	3.0 V	100 mA
Fixed Sources			
Outputs	IC		

Table 6.1 SMU set up for the measurement of R_E on the lateral DUT.

Setup Name : Setup # 2 RC			
Function	MAIN	STEP	CONSTANT
Source Name	VC	IB	VE
Sweep Mode	LIN	LIN	
Start	-500 mV	-10 mA	0.0 V
Stop	-600 mV	-10.3 mA	
# of Points	11	4	
Compliance	100 mA	3.0 V	100 mA
Fixed Sources			
Outputs	IC		

Table 6.2 SMU set up for the measurement of R_C on the lateral DUT.

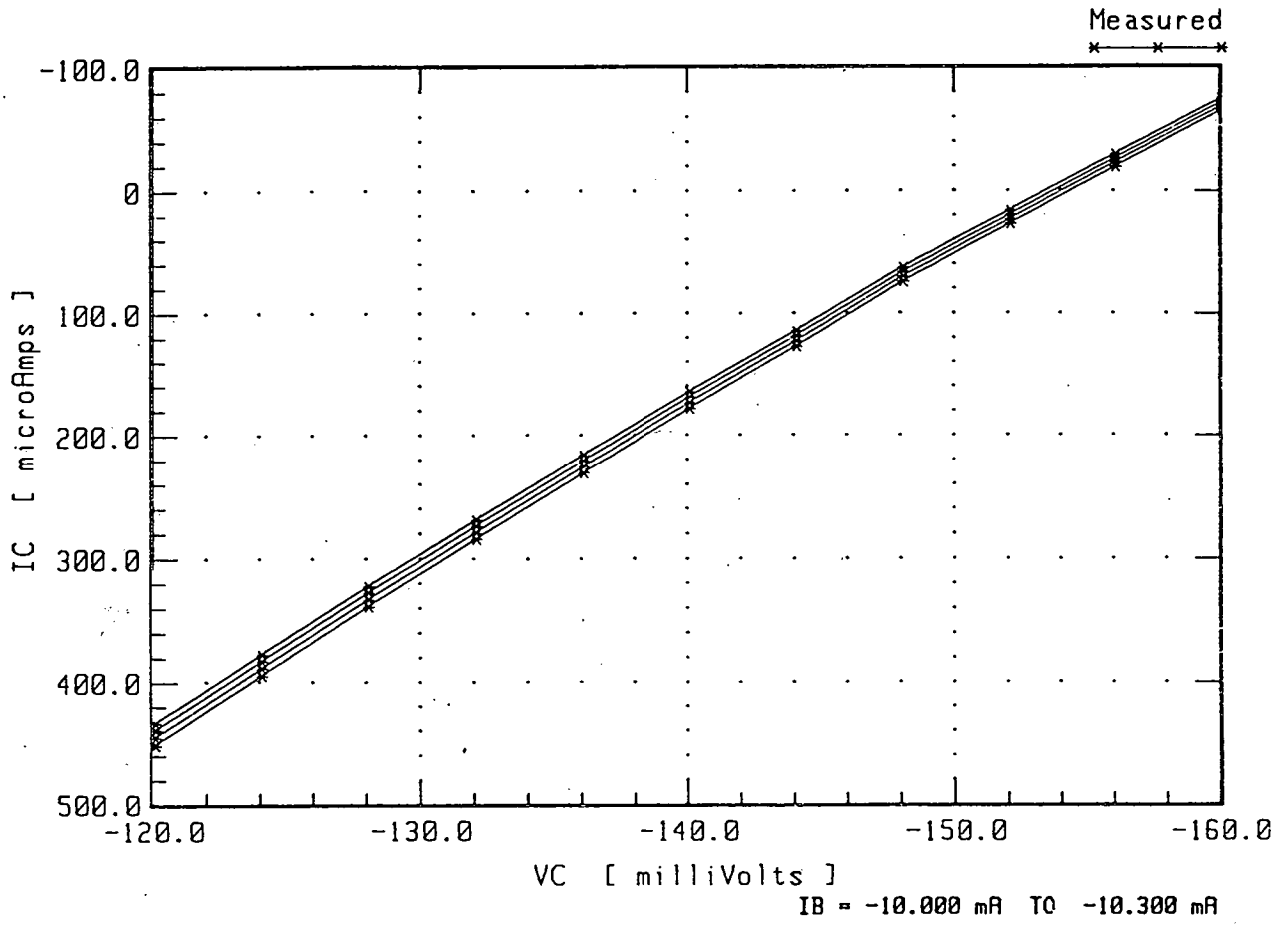


Figure 6.11 Measured curves extracted from the DUT for R_E extraction. Extracted from device type illustrated in figure 6.8b.

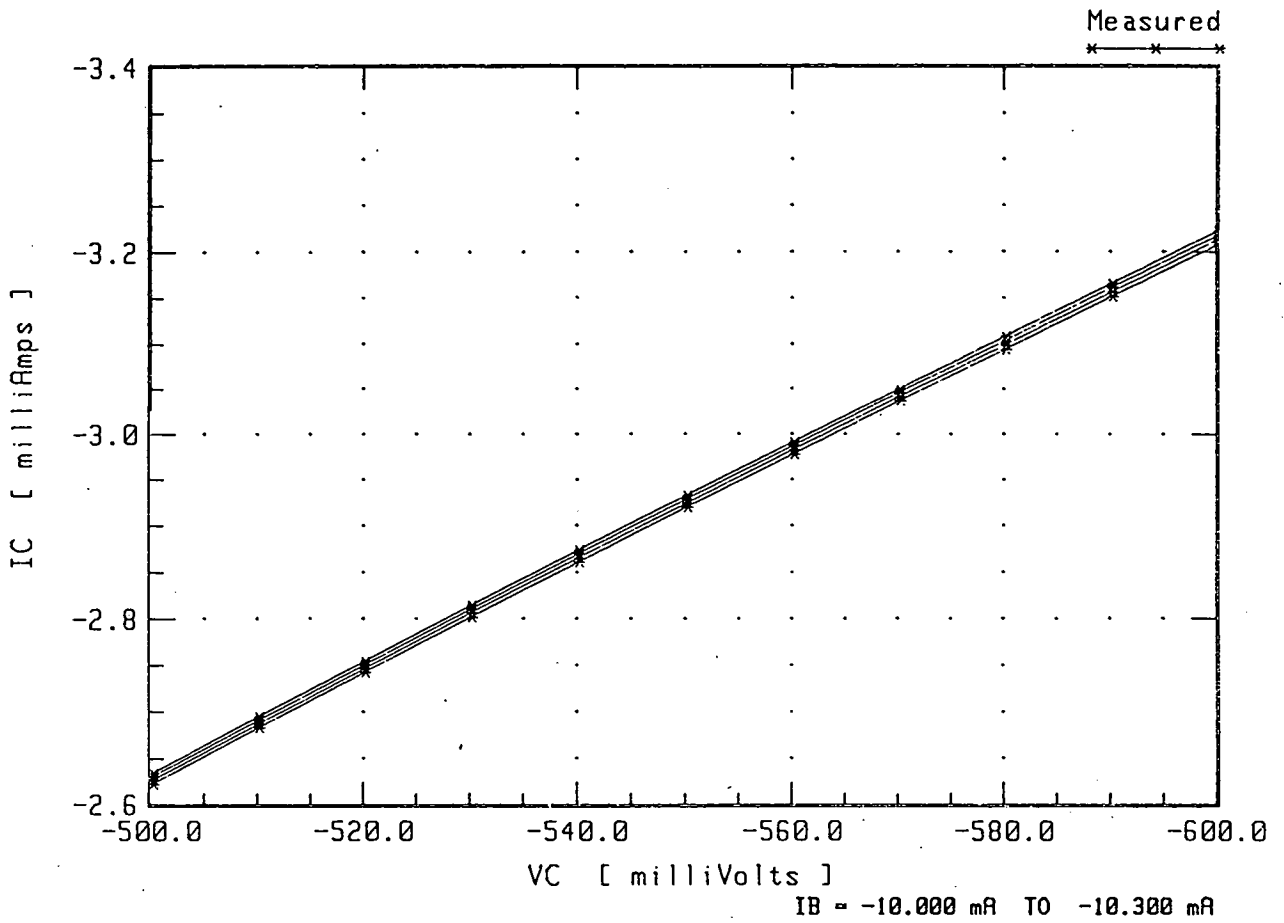


Figure 6.12 Measured curves extracted from the DUT for R_C extraction. Extracted from device type illustrated in figure 6.8b.

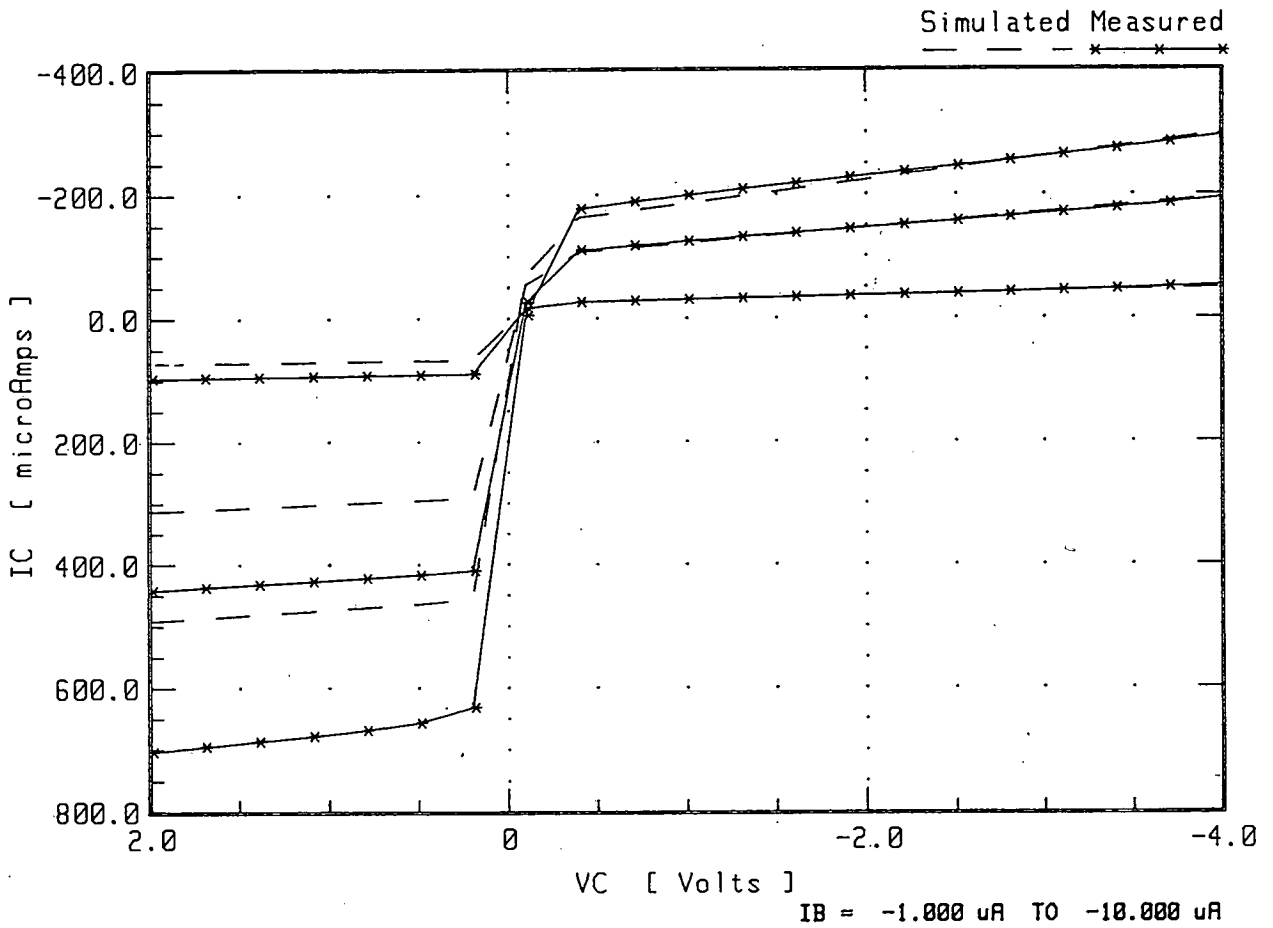


Figure 6.13 I_C vs $\pm$ V_{CE} at I_B, showing region of extraction of V_{AF} and V_{AR} the forward and reverse early voltage.

Extracted from device type illustrated in figure 6.8b.

Setup Name : Setup # 3 Early Voltage			
Function	MAIN	STEP	CONSTANT
Source Name	VC	IB	VE
Sweep Mode	LIN	LIN	
Start	2V	-1μ A	0.0 V
Stop	-4V	-10μ A	
# of Points	20	3	
Compliance	100 mA	3.0 V	100 mA
Fixed Sources			
Outputs	IC		

Table 6.3 SMU set up for the measurement of V_{AF} and V_{AR} on the lateral DUT.

Setup Name : Setup # 4 IcIb vs Vb			
Function	MAIN	CONSTANT	CONSTANT
Source Name	VB	VC	VE
Sweep Mode	LIN		
Start	-380 mV	-5 V	0.0 V
Stop	-900 mV		
# of Points	22		
Compliance	100 mA	100 mA	100 mA
Fixed Sources			
Outputs	IB	IC	

Table 6.4 SMU set up for the measurement of the forward Gummel parameters on the lateral DUT.

in chapters 2 and 4. These parameters are extracted from a single set of measured curves covering the forward and reverse areas of operation. The region, I_c vs $\pm V_{CE}$ at I_B, is shown in figure 6.13. Values of V_{AF} and V_{AR} are found by extrapolating the linear region of the curves to the x axis for both forward and reverse operation. Values extracted for the DUT were V_{AF} = 6.05V and V_{AR} = 33.3V.

6.6.3 Forward Gummel Characterisation

To extract the forward Gummel parameters I_s, N_F, R_B, β_F, I_{SE}, N_E and I_{KF}, the transistor base and collector current characteristics must be measured for varying V_{EB}. Figure 6.14 shows the ideal regions from which the parameters should be extracted.

Figure 6.15 shows the measured characteristic curves for the DUT and table 6.3 details the SMU set ups for the measurement. The extracted parameters are given in table 6.5.

Parameter	Value
I_s	472 aA
N_F	1.046
R_B	1.162 kΩ
β_F	17.63
I_{SE}	21.38 aA
N_E	1.039
I_{KF}	242.5 μA

Table 6.5 Extracted forward Gummel parameters.

6.6.4 Reverse Gummel Extraction

In reverse mode the emitter of the transistor effectively becomes the collector. In the case of the lateral bipolar transistor the forward and the reverse characteristics are very similar. This is because the emitter and the collector have a similar structure. The twin bar device has a collector and emitter which are identical (Chapter 4). Figure 6.16 shows

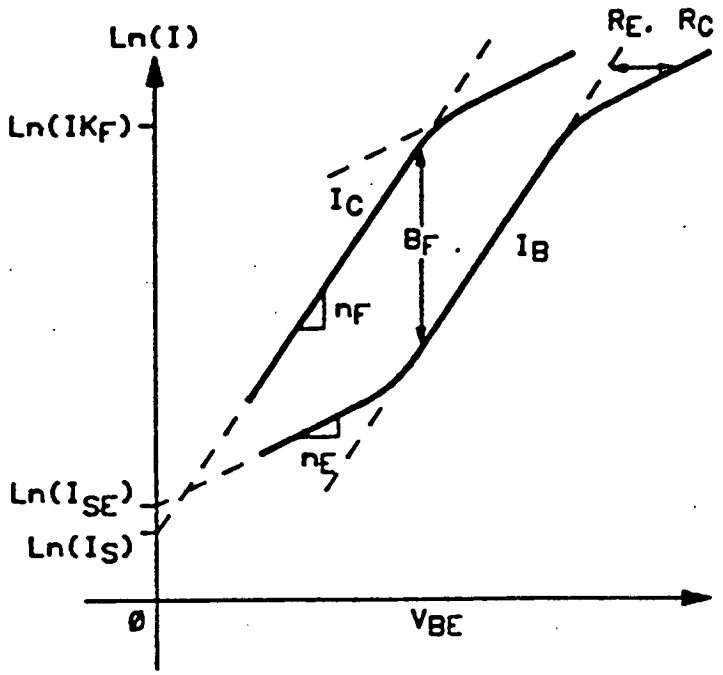


Figure 6.14 Ideal extraction regions for the Gummel parameters I_S, N_F, R_B, β_F, I_{SE}, N_E and I_{KF}.

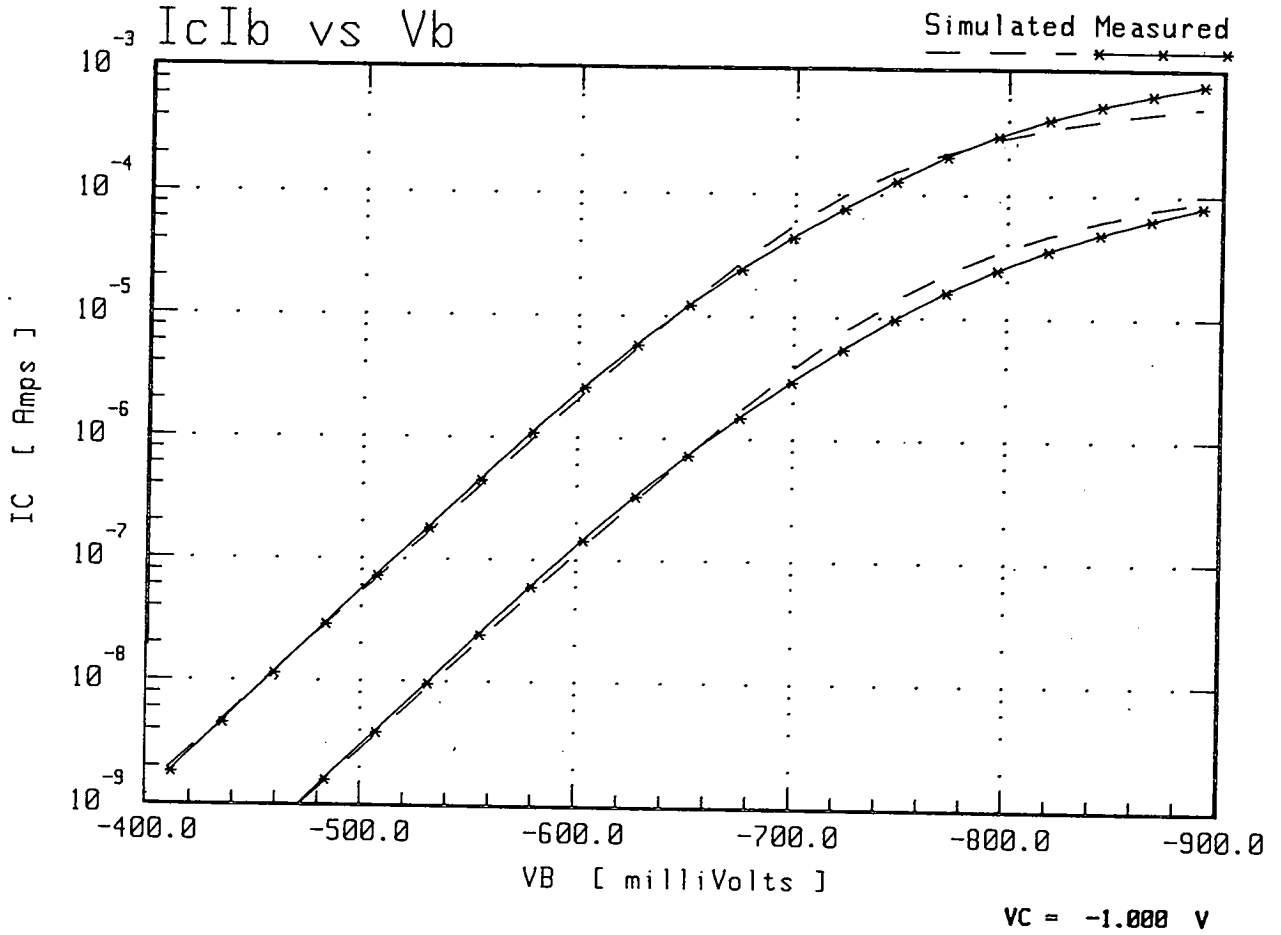


Figure 6.15 Measured forward Gummel curves for the DUT. Extracted from device type illustrated in figure 6.8b.

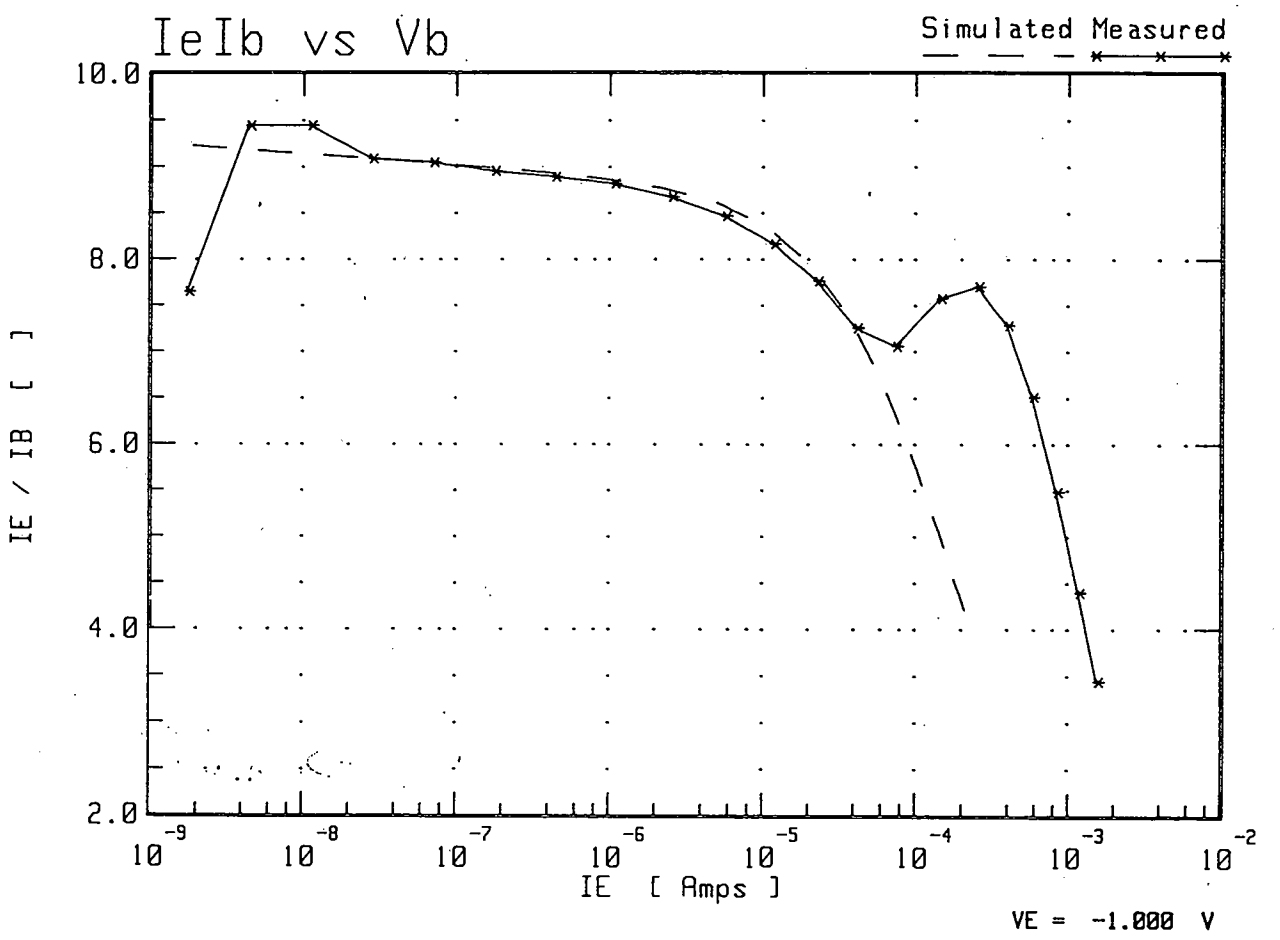


Figure 6.16 Measured reverse gain curve for the DUT.
Extracted from device type illustrated in figure 6.8b.

the measured reverse gain curve for the DUT and table 6.4 details the SMU set-ups for the measurement.

6.6.5 Parameter Optimisation

A parameter optimisation routine based on a non linear least square fit algorithm is included in the program TECAP. Parameters selected are re-calculated until the observed RMS error between the measured data and the simulated data reaches a minimum. To re-calculate the parameters, TECAP uses the Levenberg-Marquardt algorithm. This combines the steepest decent method and the Gauss - Newton method. Figure 6.17 is a flow chart describing this process.

The optimiser functions by first simulating the data using the selected model and the available parameters. These parameters may be those extracted from the measured data curves or a previously optimised set. Parameters may also be entered from the keyboard by the user. This step is the most time-consuming in the optimisation process.

In order to reduce optimisation time, good initial parameters are important. The least number of points which characterise the curve should also be chosen. By keeping the number of parameters optimised at the same time to a minimum, the optimisation time is reduced. This last constraint also serves to isolate parameter interactions, ideally each parameter should be optimised individually on a part of the device's curve which was solely or predominantly dependent on that parameter. This is not feasible as very rarely does only one parameter model a device for particular operating conditions. Hence, the user needs to understand details of the model equations, how model parameters effect model characteristics and how model parameters interact. The knowledge of this interaction is very important as there are more than one set of parameters which can produce identical characteristics.

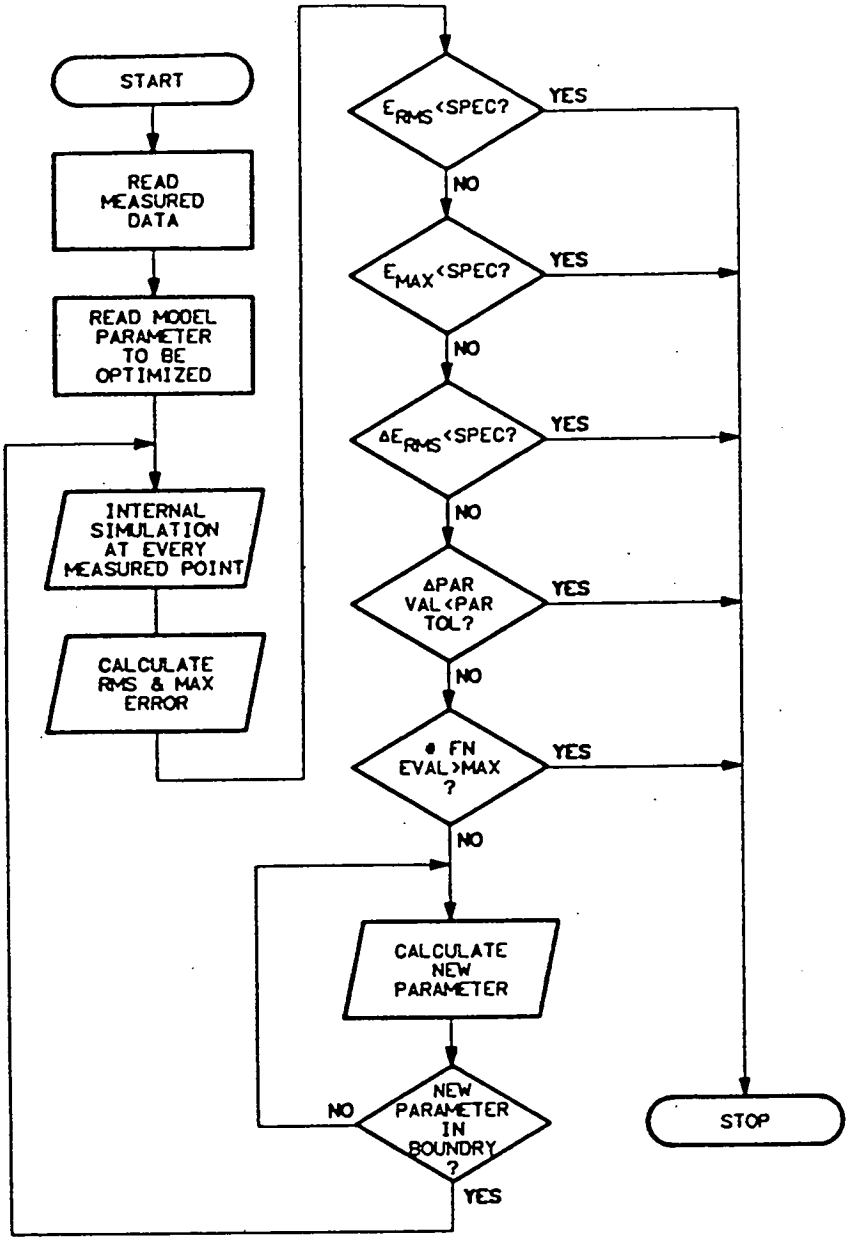


Figure 6.17 Flow chart showing the method of the Levenberg-Marquardt algorithm for the optimisation of extracted parameters.

6.6.6 Optimisation and Process Control.

Process control by device parameter extraction depends on the process variations present from die to die or wafer to wafer being highlighted by parameter variation on the same scale. The relationship between the process deviation and the parameter variation is not always straightforward and more likely than not unknown. Optimisation of the parameters in this case may lead to an improved fit of simulated to extracted values. However, the relationship between parameter and process variation may be obscured. For this reason all parameters presented in the following experiments are unoptimised unless indicated otherwise.

6.6.7 Structural Considerations

Figure 6.2 shows a schematic of the twin bar collector lateral transistor used for this experiment. It can be seen from this that there are in fact two transistors in operation, the lateral transistor of interest and a vertical transistor.

The lateral gains found on the transistors characterised ranged from 10 to 40 for the drawn base widths 3.5μm to 2μm. Figure 6.18 shows I_c plotted against V_{EB} for a typical twin bar lateral structure of drawn width 3μm. Figure 6.19 shows the lateral gain plotted against I_c. The extracted parameters are given in table 6.5. The response of the transistor is relatively flat with a gain of about 18 for collector current of 10nA to 10μA. The extracted parameters model the transistor reasonably well. The model starts to fail at higher I_c operation. The high level injection characteristics of these transistors are non-typical. The collector resistances associated with these transistors are very high compared with more conventional transistors (around 240Ω). The values of I_c used in this experiment were chosen to be in the stable (linear) region of operation.

Figure 6.20 shows the vertical gain curve associated with the same transistor used in Figure 6.19. It can be seen from this that the vertical gain is about 6 times greater than that of the lateral transistor. The properties of the vertical transistor are discussed in detail in chapter 7.

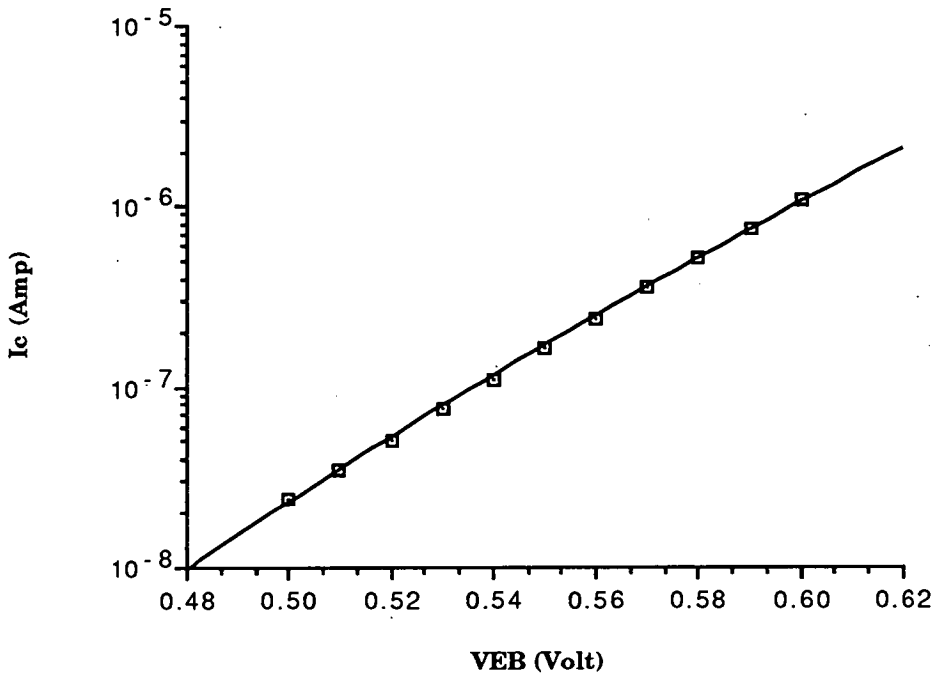


Figure 6.18 I_C plotted against V_{EB} for a typical twin bar collector structure with drawn base width = 3μm.

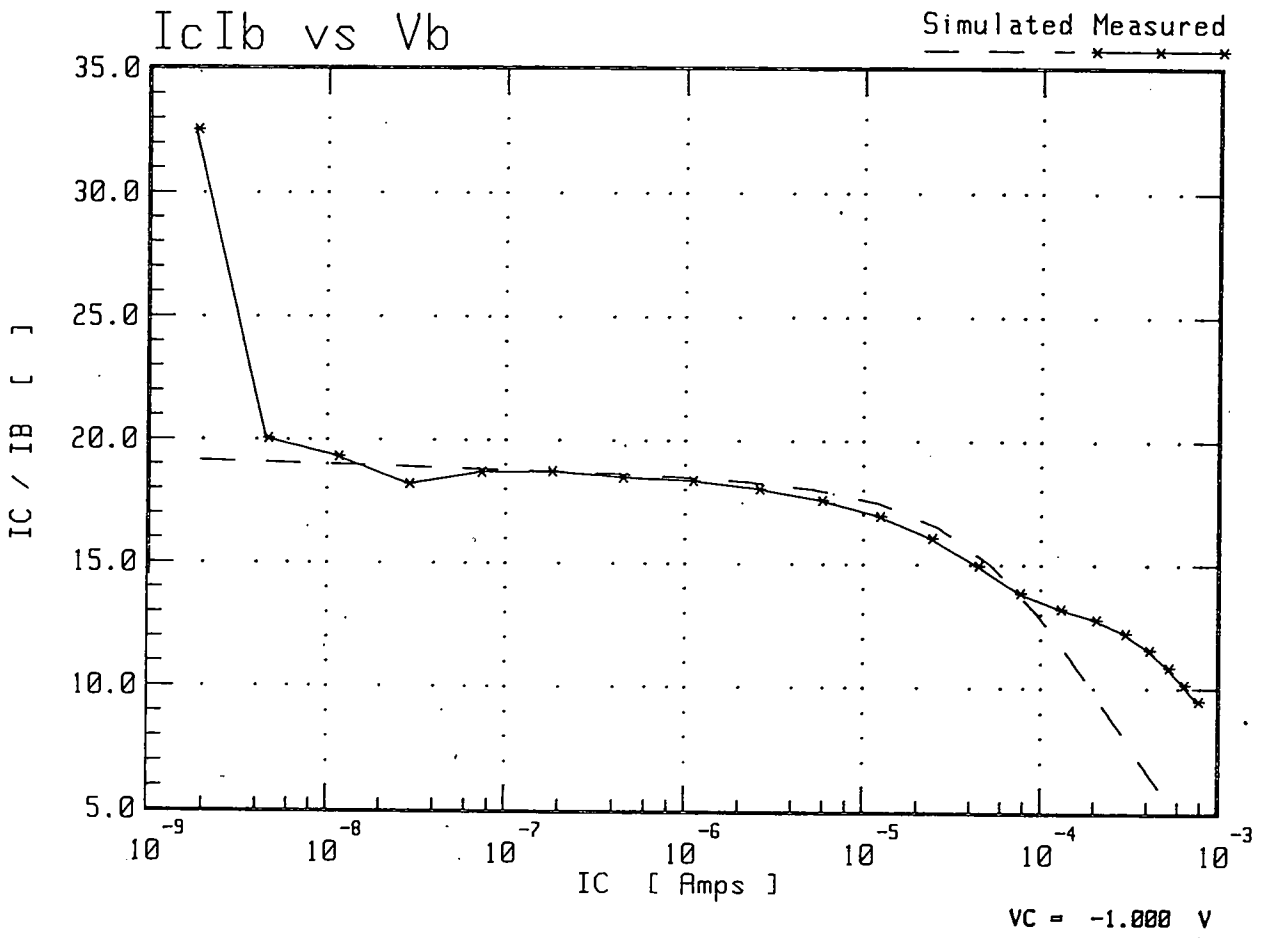


Figure 6.19 Lateral gain of a typical twin bar collector structure with drawn base width = 3μm.

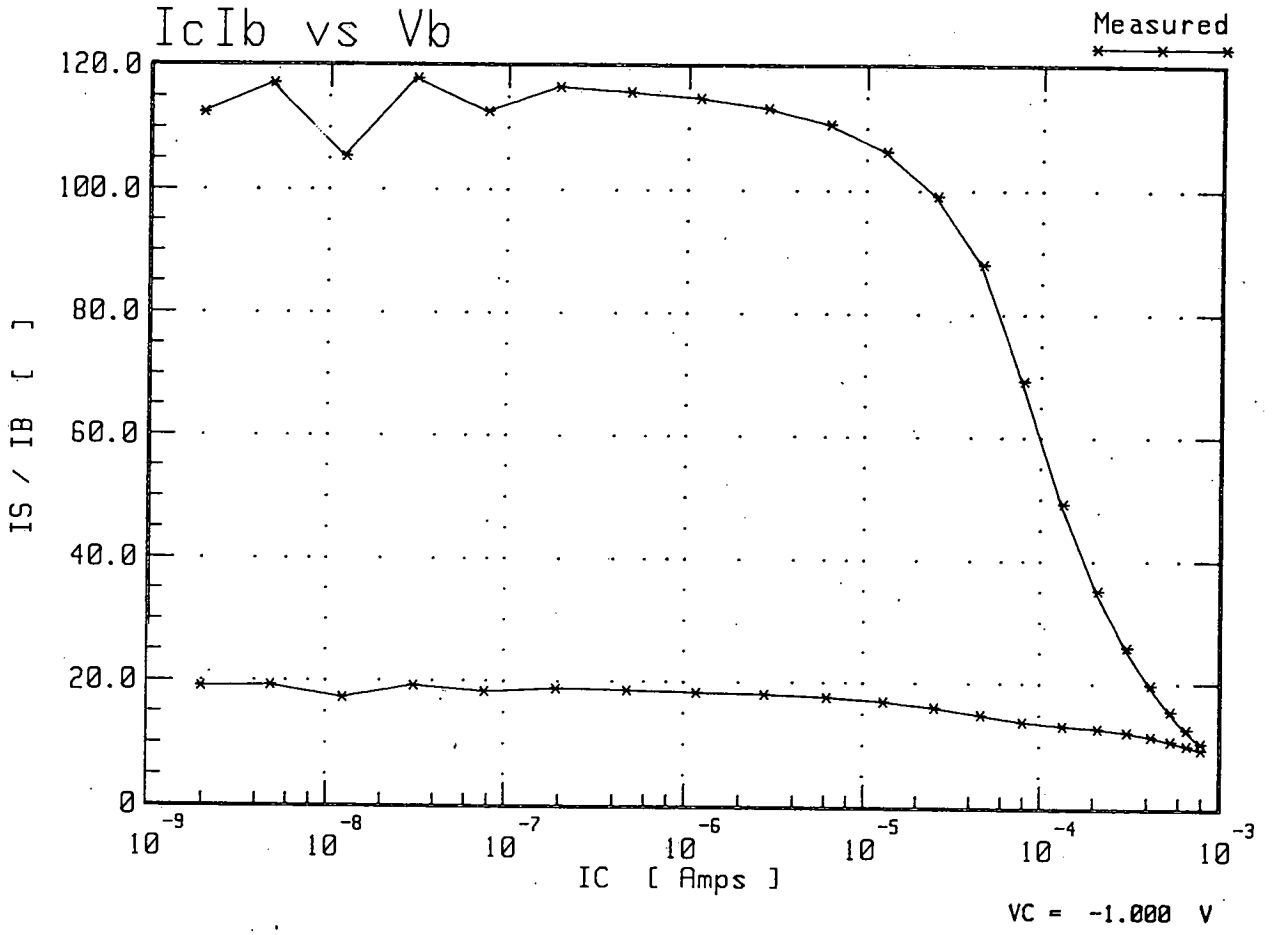


Figure 6.20 Vertical gain of a typical twin bar collector structure with drawn base width = 3μm. The transistor although primarily a lateral device has a vertical component with the well as a base and the substrate as a collector.

The emitter efficiency of the lateral transistors was poor due to the poor emitter-base doping ratio (chapter 2). For this experiment, however, the region of operation of the transistors is very restricted to maximise the effect of geometrical factors and minimise the effect of emitter-collector resistances and other doping concentration effects. These aspects have more influence on the behaviour of the device under high-level injection. High-level injection is avoided in this technique.

6.6.8 ΔL Extraction

Figure 6.21 shows the extracted $|1/I_c|$ vs drawn base width for the twin bar lateral structure shown in figure 6.8b. The intersection of the x axis is 1.64μm. This results in a value for ΔL of 0.82μm. Figure 6.22 shows $1/I_c$ vs drawn base width for the corresponding npn structure. The intersection in this case is 1.68μm giving a ΔL value of 0.84μm. The difference in ΔL measurements is due to the combination of different diffusion rates of phosphorus vs boron (npn vs pnp) and the substrate vs well doping concentration (npn vs pnp).

This method of ΔL extraction was very reliable and repeatable (complete wafer maps are shown in chapter 7). Table 6.6 shows a ten point sample of measurements across the wafer for the pnp twin bar collector 6.8b and the single bar structure 6.8a. A correlation coefficient of 0.98 means that there is no significant difference between these samples indicating that both structures perform similarly in extracting ΔL. Thus a very fast three point measurement technique was developed for extracting ΔL.

6.6.9 Simulated vs Measured Results

To understand the operation of this structure, a model was developed in the previous sections. By comparison of simulated vs measured results a qualitative feel for the accuracy of this model can be achieved.

Equation 6.15 will give a value of I_{cl} for lateral transistors. The equation needs inputs of K, x_j, W_{bo} and ΔL. The values of W_{bo} and ΔL are available from the measured results. K is known from equations 6.6 and 6.11 to be dependant on V_{EB}.

$$K = qp'_{ne}D_P l_e \tag{6.11}$$

where p'_{ne} is given by

$$p'_{ne} = \frac{n_i^2}{N_d}\left[exp\left(\frac{qV_{EB}}{kT}\right)-1\right]. \tag{6.6}$$

If D_P and N_d are assumed to be constant then K will be a constant for transistors with different drawn W_{bo} values under identical bias conditions. K will be a multiplier which affects the magnitude of the current predicted by equation 6.15. K will have no effect on the predicted intersection. With ΔL fixed by the measured results, x_j can be optimised to make the predicted intersection coincide with the measured intersection.

A routine was written in RS1 (a data analysis package) which took as inputs the measured collector current from the three lateral bipolar transistors for a fixed V_{EB}. The program then optimised x_j and K to give minimum relative error in measured vs simulated results. Figure 6.23 shows measured vs simulated results for different values of K. The measured data is from the same set of transistors as those used for figure 6.21. Table 6.7 shows the optimum value of K and x_j for ten points on the wafer with associated ΔL values and relative error in measured vs simulated. Estimated values of x_j range from 0.82μm to 0.93μm across the wafer. SUPREM3 predicts a junction depth of 0.94μm (figure 6.24). The simulated results consistently fit the measured results with a very small error. This would suggest that the model used for the basis of this experimental technique is sound.

6.6.10 Sensitivity to V_{EB}

To characterise the effect of V_{EB} on the extracted value of ΔL, ΔL was measured across the wafer with V_{EB} varied from 0.5V to 0.6V in 10mV steps. Figure 6.25 shows the extracted $|1/I_c|$ values plotted against drawn W_{bo} for several values of V_{EB}. Figure 6.26 shows the extracted ΔL vs V_{EB} for 8 of the 10 devices measured on the wafer. It can

	Extracted ΔL (μm)	
Transistor Position	Twin Bar Collector	Single Bar Collector
1	0.802	0.803
2	0.740	0.737
3	0.734	0.736
4	0.804	0.802
5	0.858	0.856
6	0.743	0.746
7	0.671	0.670
8	0.784	0.785
9	0.82	0.822
10	0.725	0.727

Table 6.6 Sample measurements comparing single bar structure and twin bar structure.

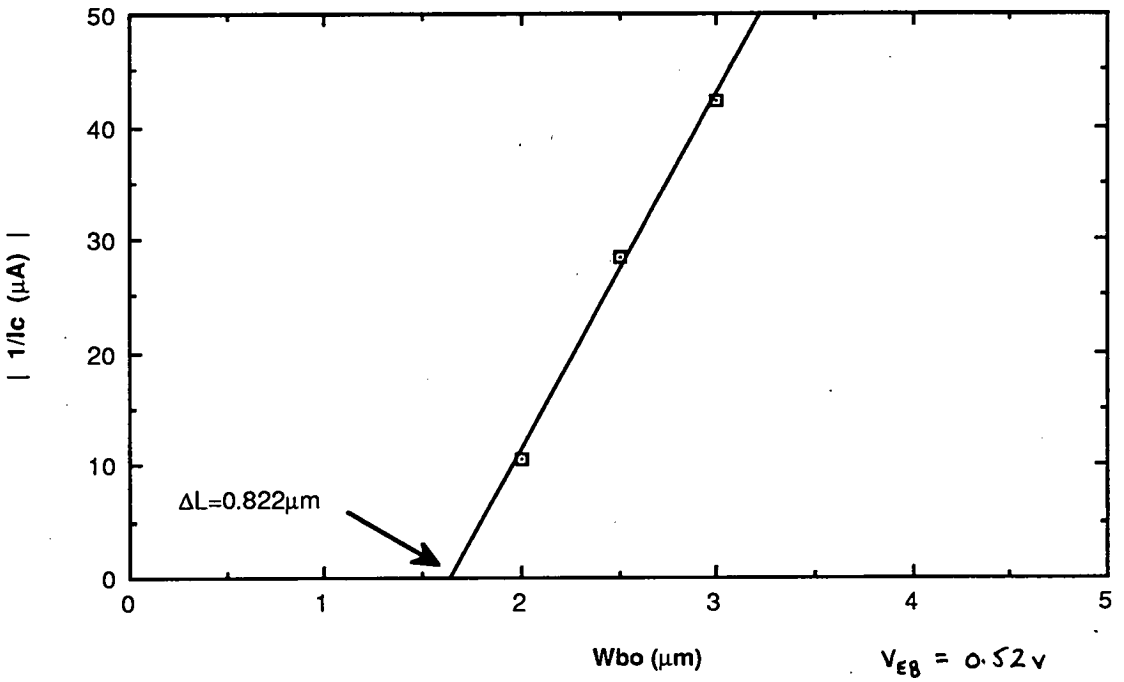


Figure 6.21 Extracted $|1/I_C|$ vs drawn base width for the twin bar collector structure shown in figure 6.8b.

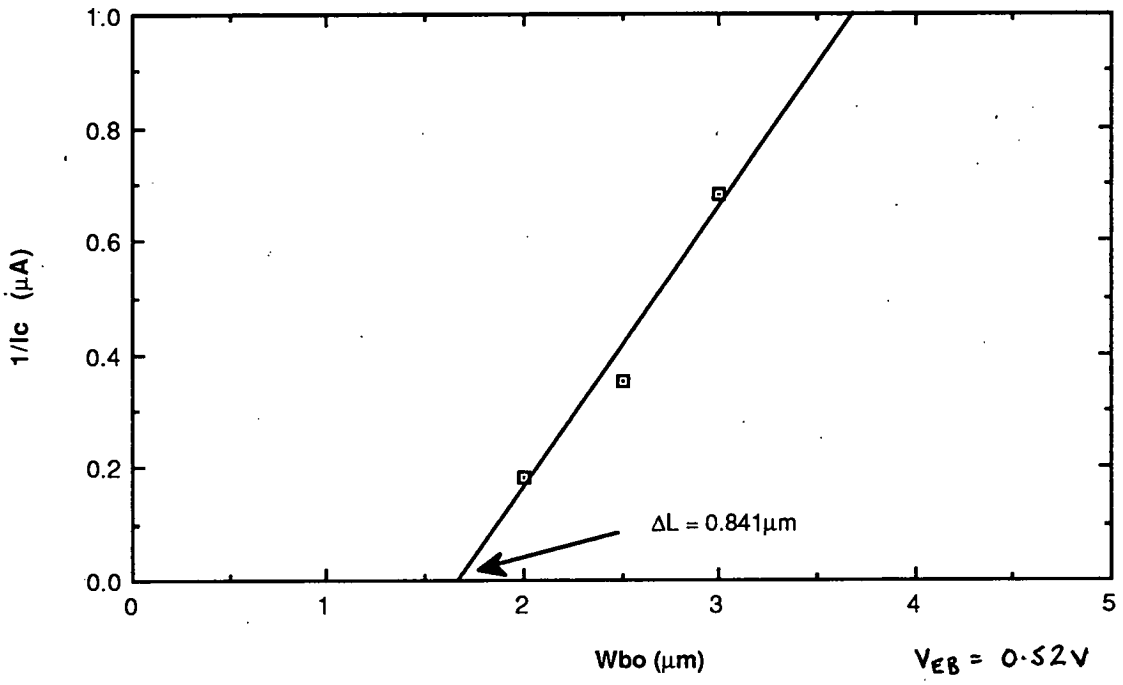


Figure 6.22 Extracted $1/I_C$ vs drawn base width for the npn twin bar collector structure.

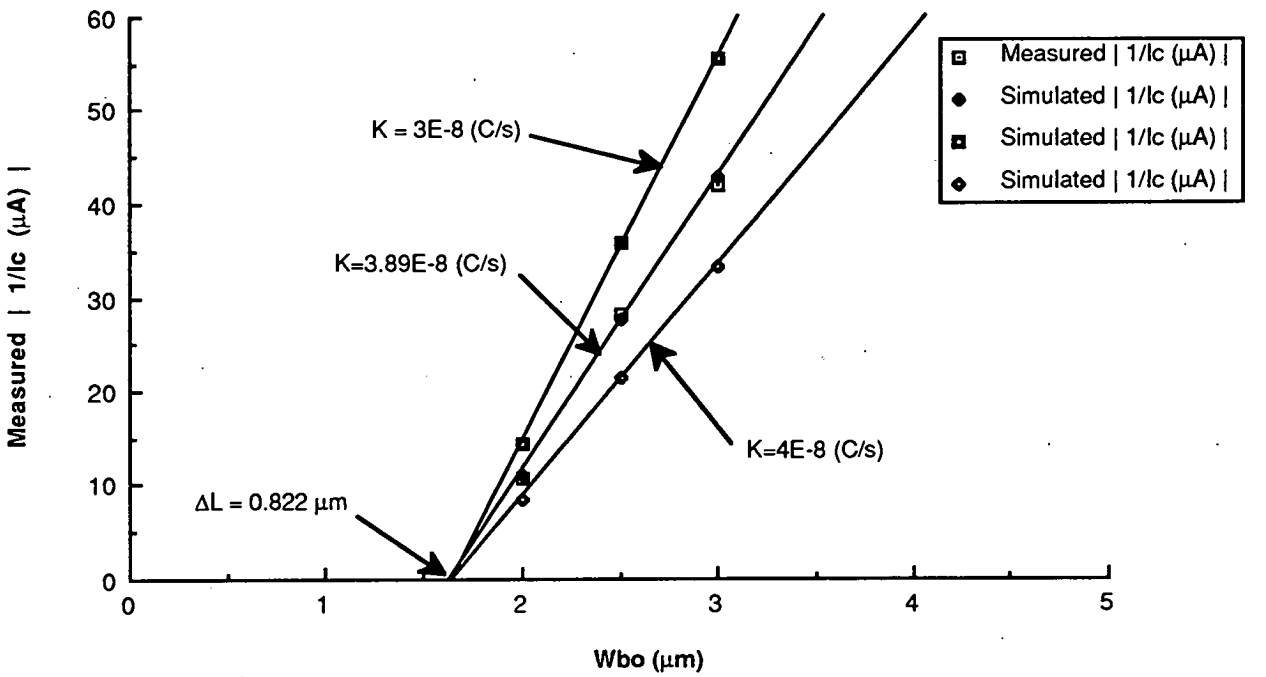


Figure 6.23 Simulated vs measured data. This illustrates close fit of simulated to measured data.

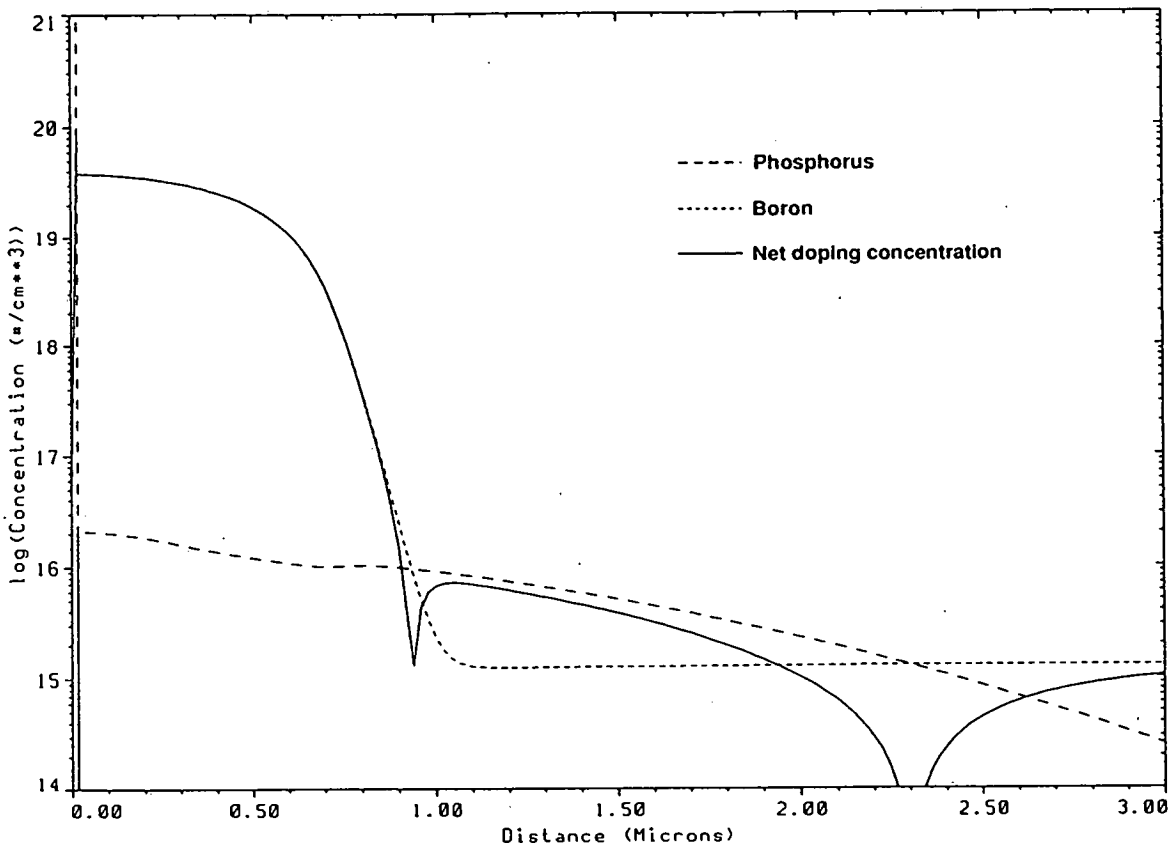


Figure 6.24 SUPREM output for a simulation of the EMF 5μm n-well CMOS process, estimating the source/drain junction depth (x_j). Not taken from the process split.

Measured ΔL (μm)	Simulated ΔL (μm)	Optimised K (Cs^{-1})	Optimised x_j (μm)	% relative error (predicted vs measured I_{cl})
0.813	0.814	4.38 E-8	0.895	7.5
0.745	0.746	5.28 E-8	0.835	4.4
0.748	0.751	6.59 E-8	0.84	1.9
0.801	0.803	5.51 E-8	0.885	1.5
0.861	0.862	5.01 E-8	0.933	1.4
0.726	0.730	5.51 E-8	0.82	2.3
0.785	0.786	4.34 E-8	0.871	3.5
0.782	0785	6.41 E-8	0.87	2.0
0.822	0.823	3.89 E-8	0.902	2.8
0.82	0.822	5.77 E-8	0.901	0.5

Table 6.7 Extracted vs simulated ΔL for 10 points across the wafer. This table shows optimised K and x_j with associated error in measured vs simulated I_{Cl}.

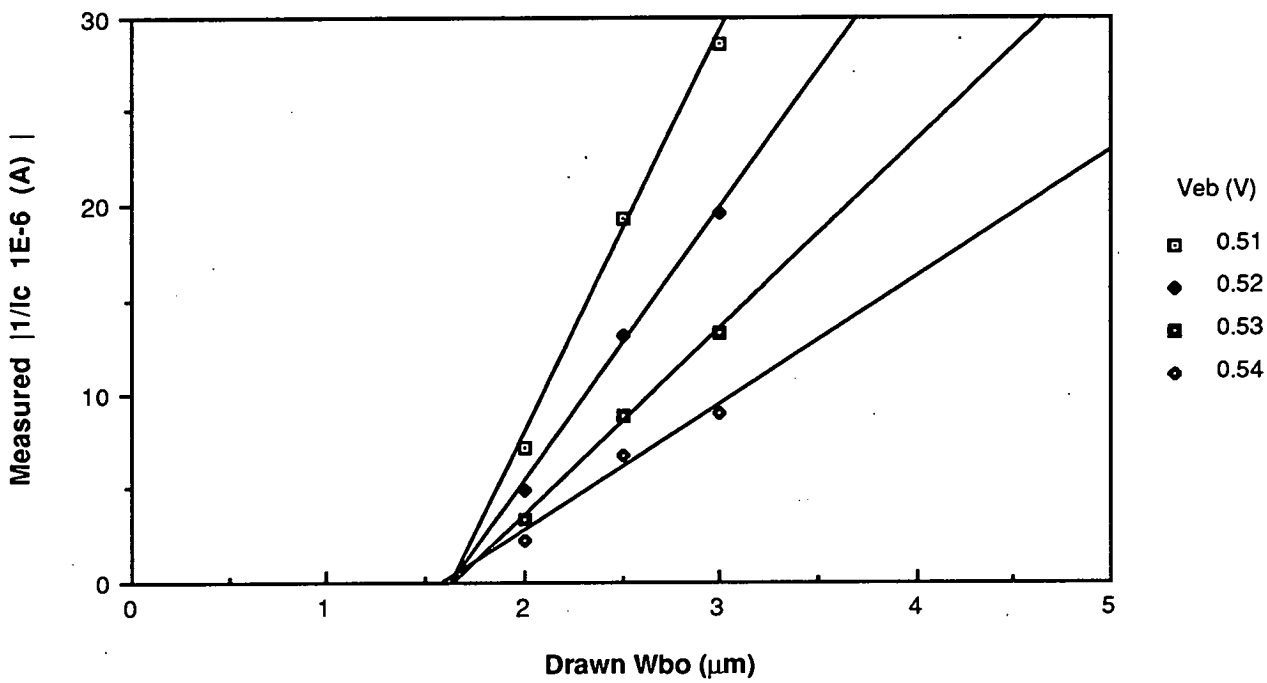


Figure 6.25 $1/I_C$ vs drawn base width plotted for several values of V_{EB}.

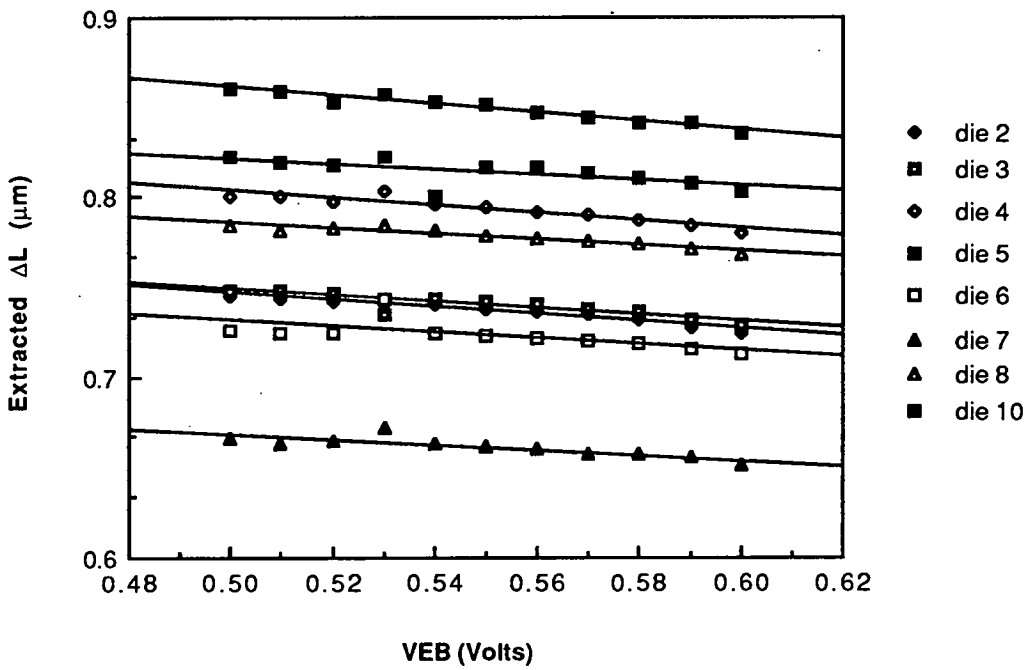


Figure 6.26 Extracted ΔL vs V_{EB} for 8 points on wafer.

be seen from this that around the specified region of operation for this method, the extracted value of ΔL is relatively insensitive to V_{EB}, the base-emitter bias conditions. Figure 6.26 shows the relationship of extracted ΔL vs V_{EB} for device 5. As V_{EB} is increased, the extracted value of ΔL remains constant until V_{EB} reaches about 0.53V then the extracted ΔL falls in a linear fashion. The drop in ΔL measured is under 2% across the experimental range. With increasing V_{EB} the base-emitter depletion region increases, the lateral transistor begins to leave the low level injection restriction imposed earlier and the relationship between I_{cl} and W_{bo} begins to fail.

6.6.11 Significance of K

In section 6.6.9 it was seen that K will be related to V_{EB} through equations 6.11 and 6.6. When K is optimised to give the best fit of the extracted data to the simulation model, the relationship of K vs V_{EB} for these transistors can be plotted. Figure 6.27 shows the relationship of K vs V_{EB} for the transistors used in Figure 6.25. From this plot p'_{ne}, the excess hole concentration, can be calculated as a function of V_{EB}. It varies from 5.3 x10^{11} cm^{-3} to 1x10^{13} cm^{-3} as V_{EB} varies from 0.5 to 0.58V.

6.6.12 Comparison With Conventional Methods

ΔL was extracted from an array of MOS test structures with varying drawn lengths. The method of ΔL extraction was described in section 6.1. Table 6.8 shows the traditional ΔL value vs that extracted from the parasitic bipolar structures on the 10 die used above. Very close agreement between the two methods was observed.

6.7 Conclusions

Lateral parasitic bipolar transistors can be used to extract accurate values of ΔL, the sideways diffusion associated with MOS transistors, quickly and efficiently. The values extracted have been shown to compare well with those extracted using conventional CMOS structures. Device biasing limits have been established and the technique found to be stable within these limits. The technique could be

used in any conventional CMOS drop-in PC or scribe grid with no extra mask sets.

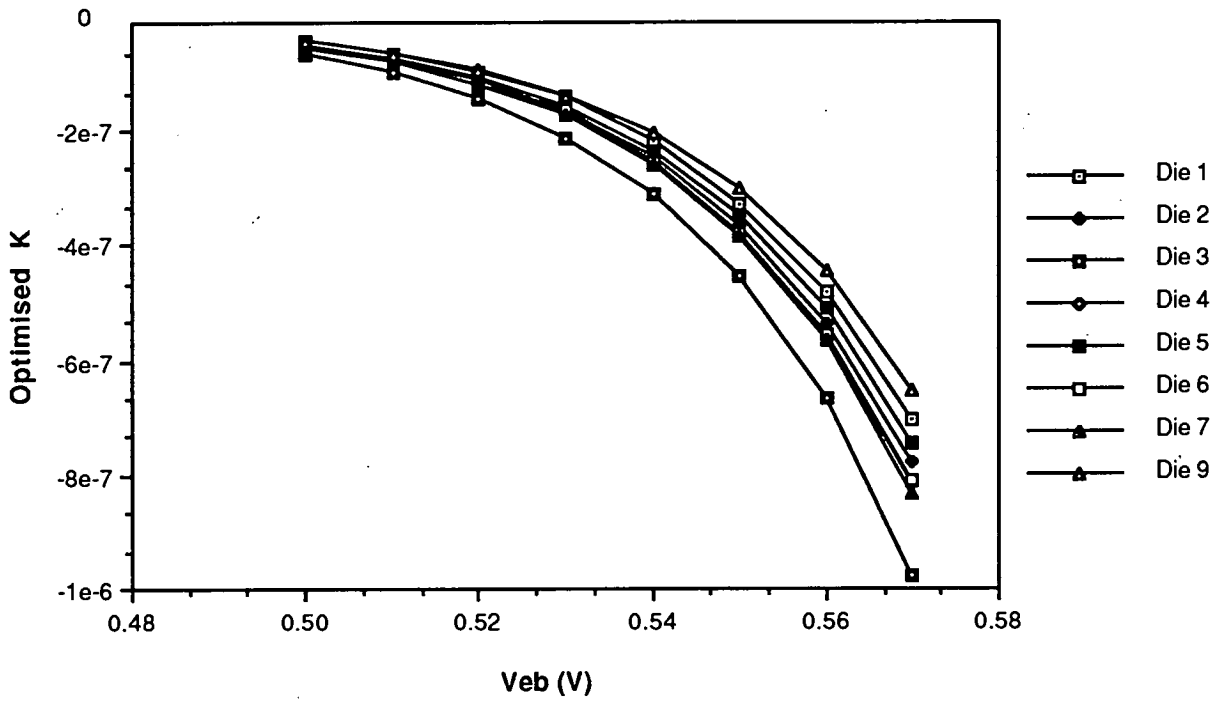


Figure 6.27 Plot showing the variation of K with V_{EB}.

	Extracted ΔL (μm)		
Die Position	Twin Bar Collector	MOS Transistors	Delta (μm)
1	0.802	0.847	0.045
2	0.740	0.703	-0.037
3	0.734	0.698	-0.036
4	0.804	0.817	0.013
5	0.858	0.809	-0.049
6	0.743	0.670	-0.073
7	0.671	0.639	0.032
8	0.784	0.827	0.043
9	0.82	0.799	-0.021
10	0.725	0.767	0.042

Table 6.8 Sample measurements comparing lateral bipolar twin bar structure and conventional MOS methods.

References

[1] D.E. Fulkerson, "A two-dimensional model for the calculation of common emitter current gains of lateral pnp transistors," *Solid-State Electron.*, Vol 11, p. 821, 1968.

[2] H.K.J. Ihantola and J.L. Moll, "Design theory of a surface Field-Effect Transistor," *Solid-State Electron.*, Vol. 7, p. 423, 1964.

[3] J. Lindemeyer and W. Schneider, "Theory of lateral Transistors," *Solid-State Electron.*, Vol. 10, p.225, 1967.

[4] S. Chou, "An Investigation of lateral transistors D.C. Characteristics," *Solid-State Electron.*, Vol. 4, p.811, 1971.

[5] K. Seo and C. Kim, "On the Geometrical Factor of Lateral pnp Transistors," *IEEE Trans. Electron. Devices*, Vol. ED-27, p. 295, 1980.

Chapter 7

CMOS Process Uniformity Evaluation Through the Characterisation of Parasitic Transistors

To obtain consistent yields on any process, uniformity of processing is of the utmost importance. Yield loss can be classified into three main groups,

1.Mis-processing

This occurs where the product misses out a step or receives the wrong step in the process. This can be controlled to an extent by tight manufacturing rules and procedures. It should be relatively rare and easy to diagnose.

2. Defectivity

Defects are particles introduced from humans, wafers (silicon debris from smashed or chipped wafers) and machinery, both through wear and tear and unexpected chemical reactions. Defectivity can be reduced through better housekeeping and machine maintenance. In-line inspections help to catch defects, at the end of the line SEM and EDAX inspections help to trace defects.

3. Process non-uniformity

Wafer to wafer, across wafer and batch to batch non-uniformity can result in more yield loss than the other two categories. Non-uniform processing is usually detected by drift in device parameters or wafer maps which reveal cross wafer variations. These techniques are used here to evaluate test structures designed to detect and monitor process uniformity.

Chapters 5 and 6 dealt with specific structures designed and characterised to provide information on traditional CMOS performance parameters, ie effective channel length and well depth. This chapter discusses and presents results on the relationship between traditional CMOS parameters and those of the parasitic transistors available from the process. Potential increases in uniformity sensitivity are highlighted. Other relationships which are interesting for their physical device aspects are also shown.

7.1 The Vertical Parasitic Bipolar Transistor

The bulk of the characterisation work presented in this chapter comes from the vertical parasitic bipolar structure. Much important work has been done on the inclusion of vertical bipolar transistors into a CMOS process [1,2,3,4,5,6,7]. The emphasis behind this work has been to develop bipolar transistors as part of a BICMOS process, or to reduce and understand the effects of bipolar transistors on the latch-up problem associated with CMOS devices. The work presented here focuses more on the ability of the parasitic vertical bipolar transistor to be used as a simple test structure for process control and uniformity evaluation.

7.1.1 Device Structure

As detailed in chapter 4, several vertical bipolar transistors were designed and fabricated. Figure 7.1 shows a schematic of a typical device fabricated in a n-well CMOS process. By examining this schematic we would expect this transistor to be sensitive to variations in n-well depth, doping profile and concentration. Since the emitter is formed by the p+ source/drain implant, the device will also be sensitive to source/drain

doping profile. We would infer then that the device might be sensitive to implant variations and back-end heat treatments.

7.1.2 Characterisation Technique

The devices were characterised using TECAP, a parametric extraction software package. The hardware requirements and capabilities of this package were discussed earlier in chapter 3. An automatic wafer prober was used to facilitate wafer mapping of parameters. This enabled cross-wafer non-uniformity to be investigated. Wafer mapping has long been known as an effective technique for this purpose [8].

7.1.3 Parametric Results

A detailed characterisation of the vertical bipolar parasitic device is presented below. The devices were fabricated on wafers which had a n-well drive-in time split. The details of this split were discussed in chapter 4. A full characterisation for each part of the n-well depth split is presented. The results are typical of those extracted across the wafer and used for process uniformity discussions later in this chapter. Results are presented for wafers 1,8,6 and 10. These wafers had drive-in times of 18,8,4 and 1 hr respectively. The transistors were characterised in common emitter mode.

7.1.4 R_E and R_C

Tables 7.1 and 7.2 show the SMU set-ups for R_E and R_C extraction. Figure 7.2a-h show a typical set of curves used to extract the parameters R_E and R_C for each part of the split. Figure 7.3a-b shows the wafer mean (65 points) of R_E and R_C vs base junction depth. Figure 7.3a shows a decrease in R_E with drive-in time. As the base (n-well) is driven in farther the effective dopant in the emitter increases and the resistance drops. Figure 7.3b shows that the collector resistance R_C of the device also drops by around 10% as the drive-in time increases. This is perhaps more difficult to explain. The collector resistance will be dominated by the contact to substrate and then the resistive path from the base to the collector contact. These should be almost identical in all the devices. The

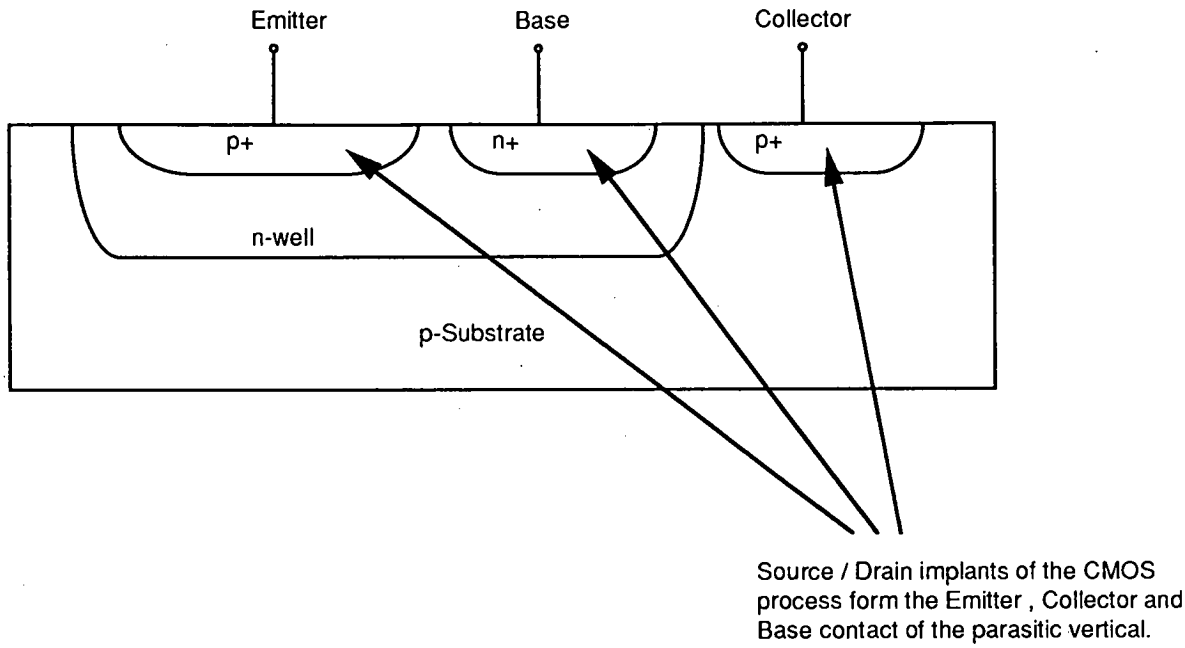


Figure 7.1 Schematic of typical parasitic vertical bipolar transistor fabricated in a n-well CMOS process.

Setup Name : RE	Setup # 1		
Function	MAIN	STEP	CONSTANT
Source Name	VC	IB	VE
Sweep Mode	LIN	LIN	
Start	-120 mV	-10 mA	0.0 V
Stop	-160 mV	-10.3 mA	
# of Points	11	4	
Compliance	100 mA	3.0 V	100 mA
Fixed Sources			
Outputs	IC		

Table 7.1 SMU set up for the measurement of R_E.

Setup Name : RC	Setup # 2		
Function	MAIN	STEP	CONSTANT
Source Name	VC	IB	VE
Sweep Mode	LIN	LIN	
Start	-500 mV	-10 mA	0.0 V
Stop	-600 mV	-10.3 mA	
# of Points	11	4	
Compliance	100 mA	3.0 V	100 mA
Fixed Sources			
Outputs	IC		

Table 7.2 SMU set up for the measurement of R_C.

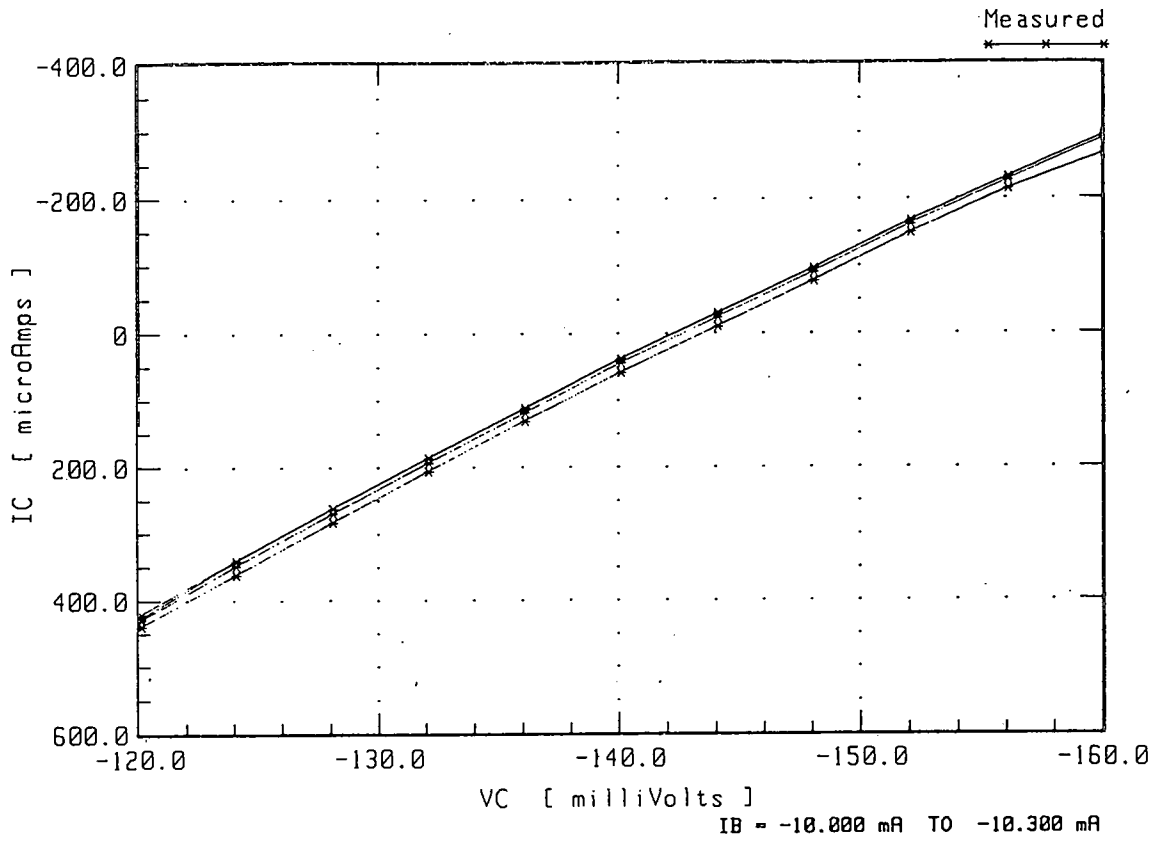


Figure 7.2a Typical measured curves for R_E extraction, taken from wafer 1.

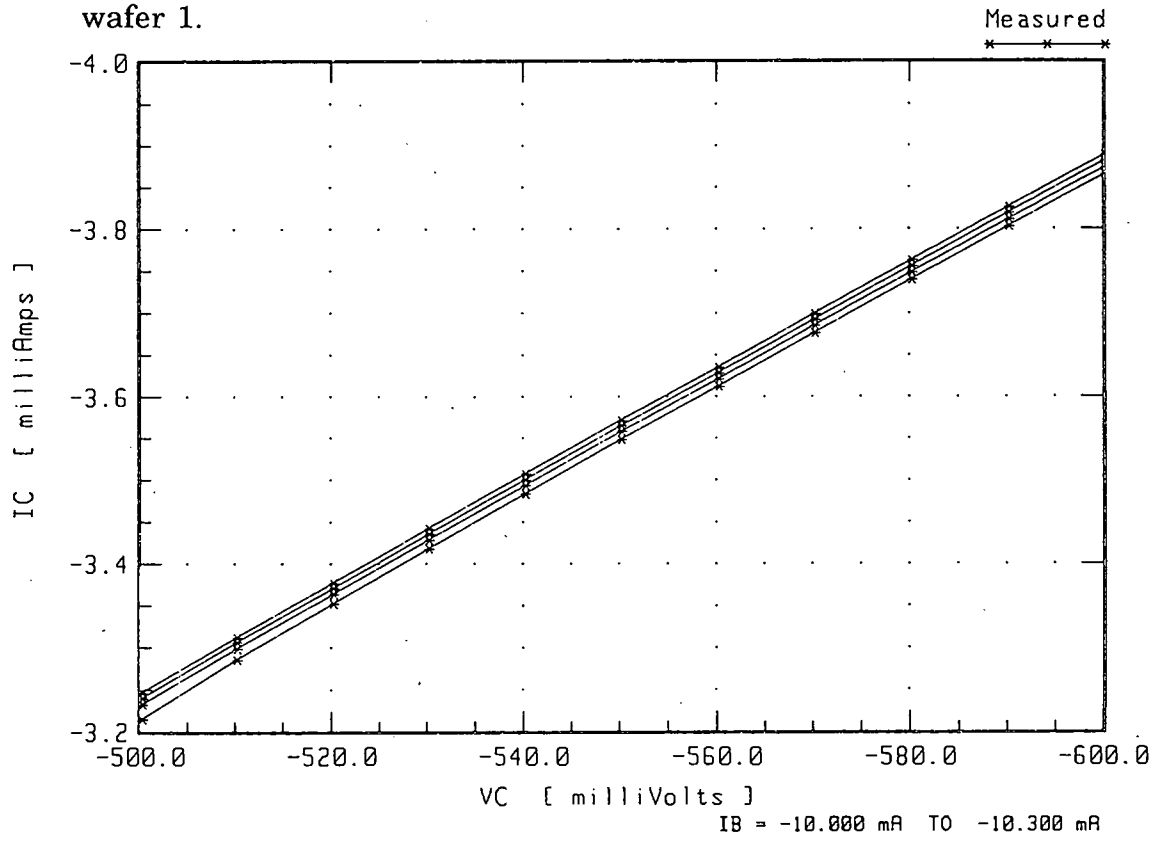


Figure 7.2b Typical measured curves for R_C extraction, taken from wafer 1.

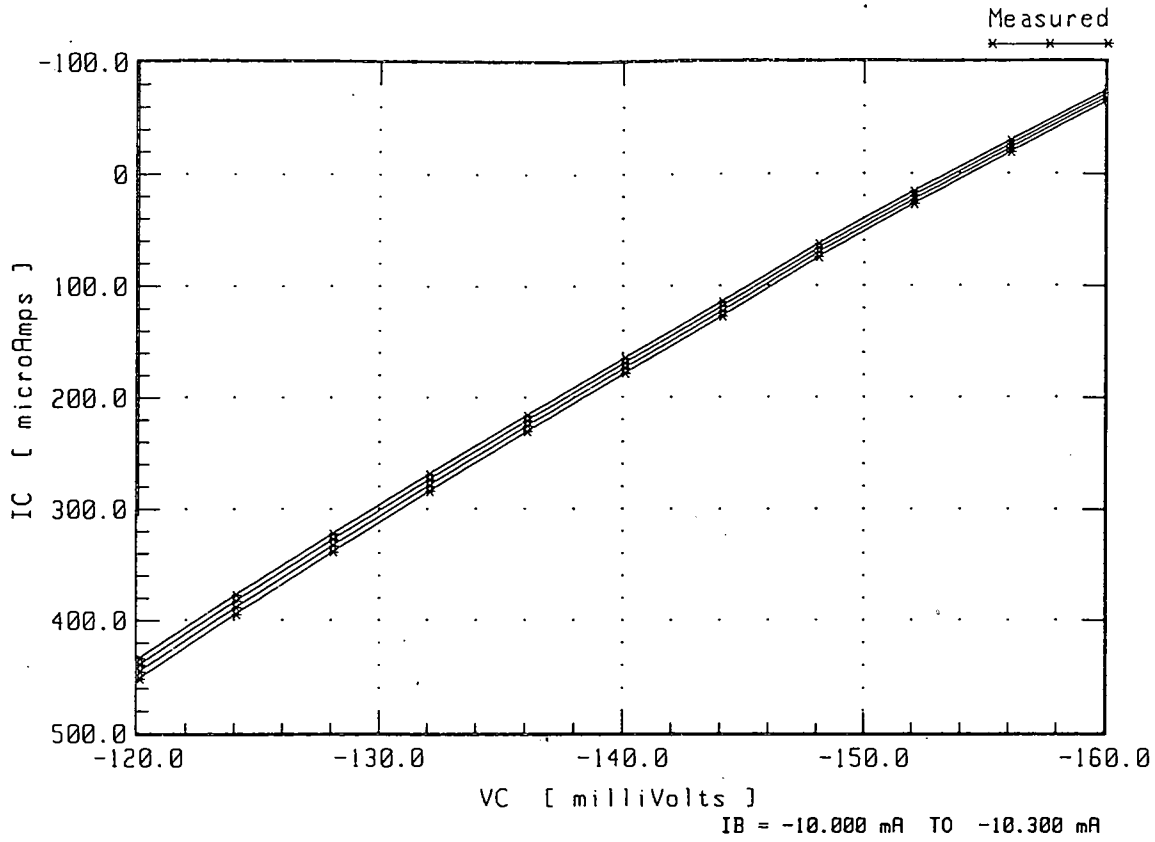


Figure 7.2c Typical measured curves for R_E extraction, taken from wafer 4.

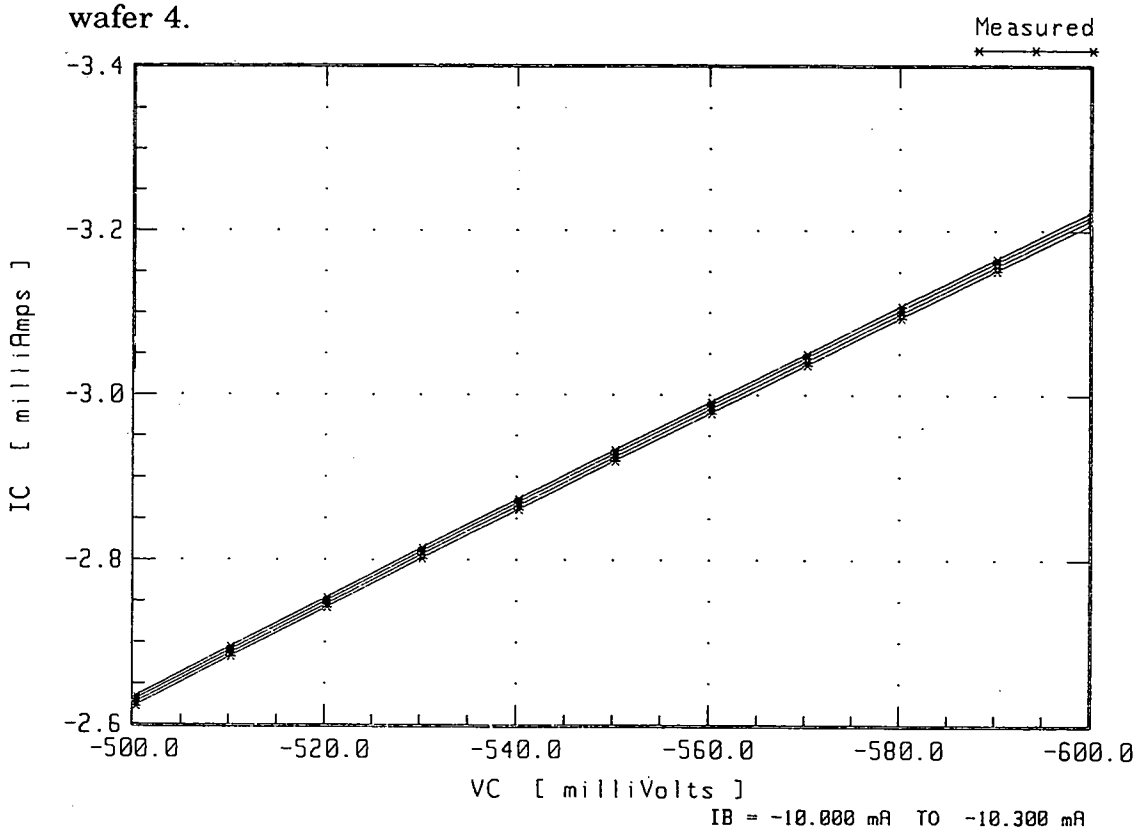


Figure 7.2d Typical measured curves for R_C extraction, taken from wafer 4.

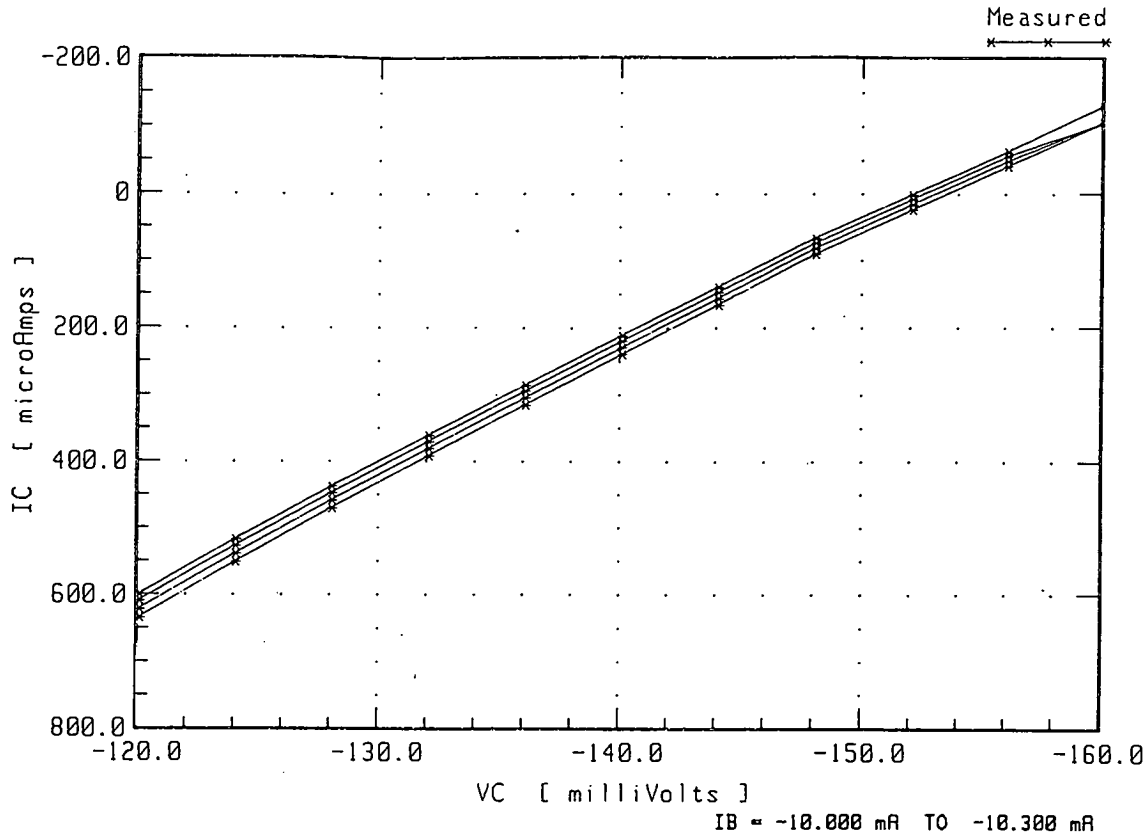


Figure 7.2e Typical measured curves for R_E extraction, taken from wafer 6.

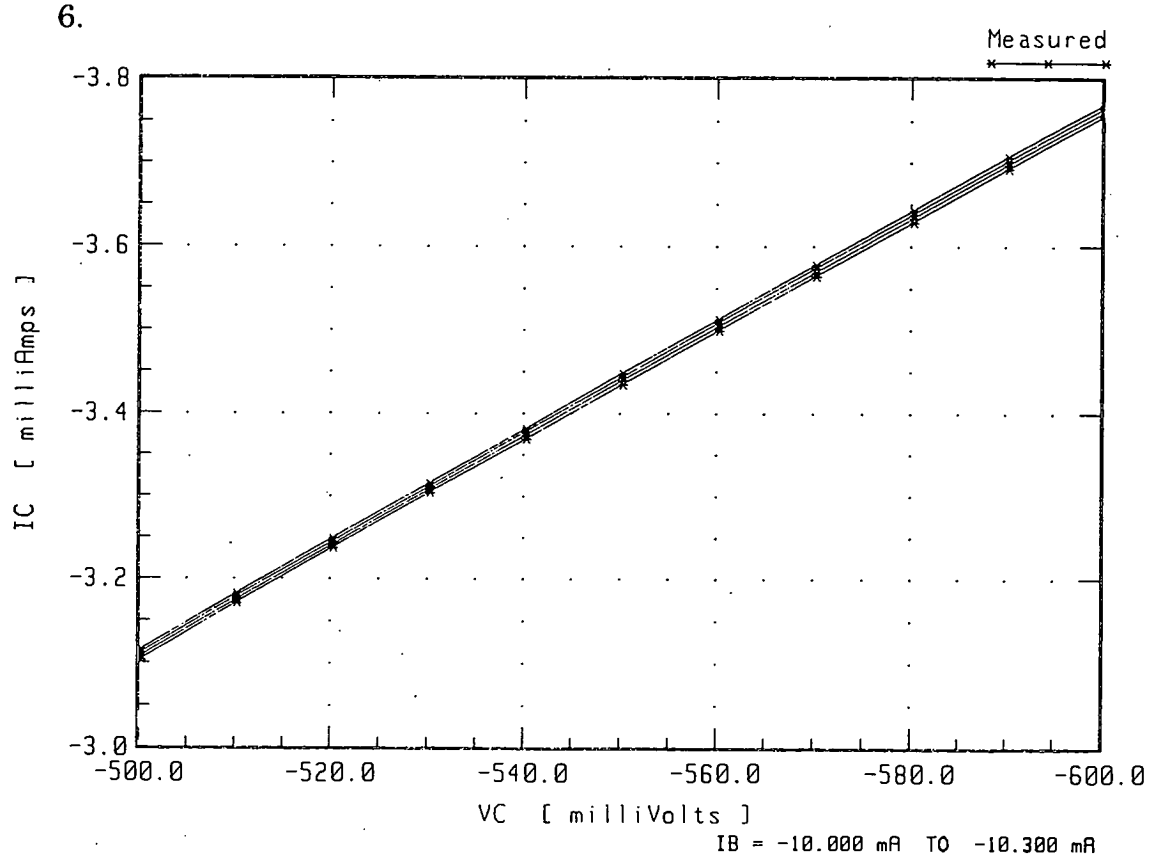


Figure 7.2f Typical measured curves for R_C extraction, taken from wafer 6.

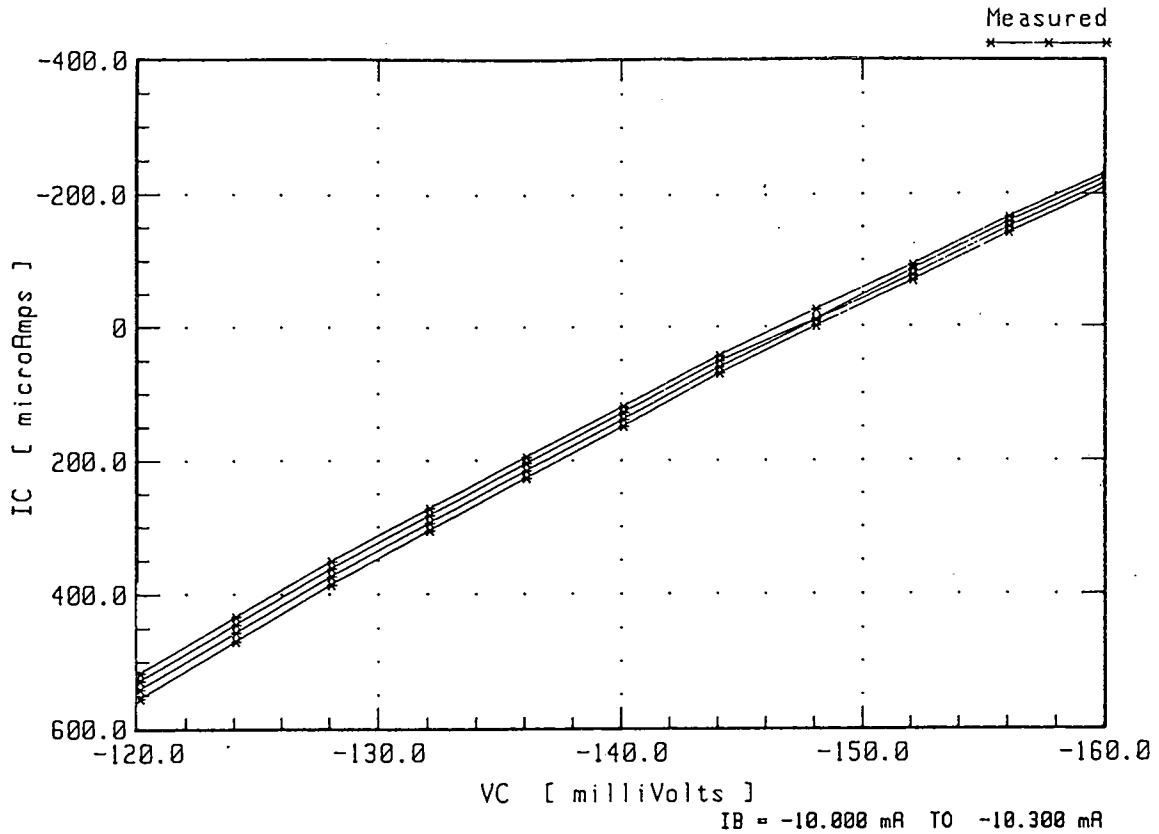


Figure 7.2g Typical measured curves for R_E extraction, taken from wafer 10.

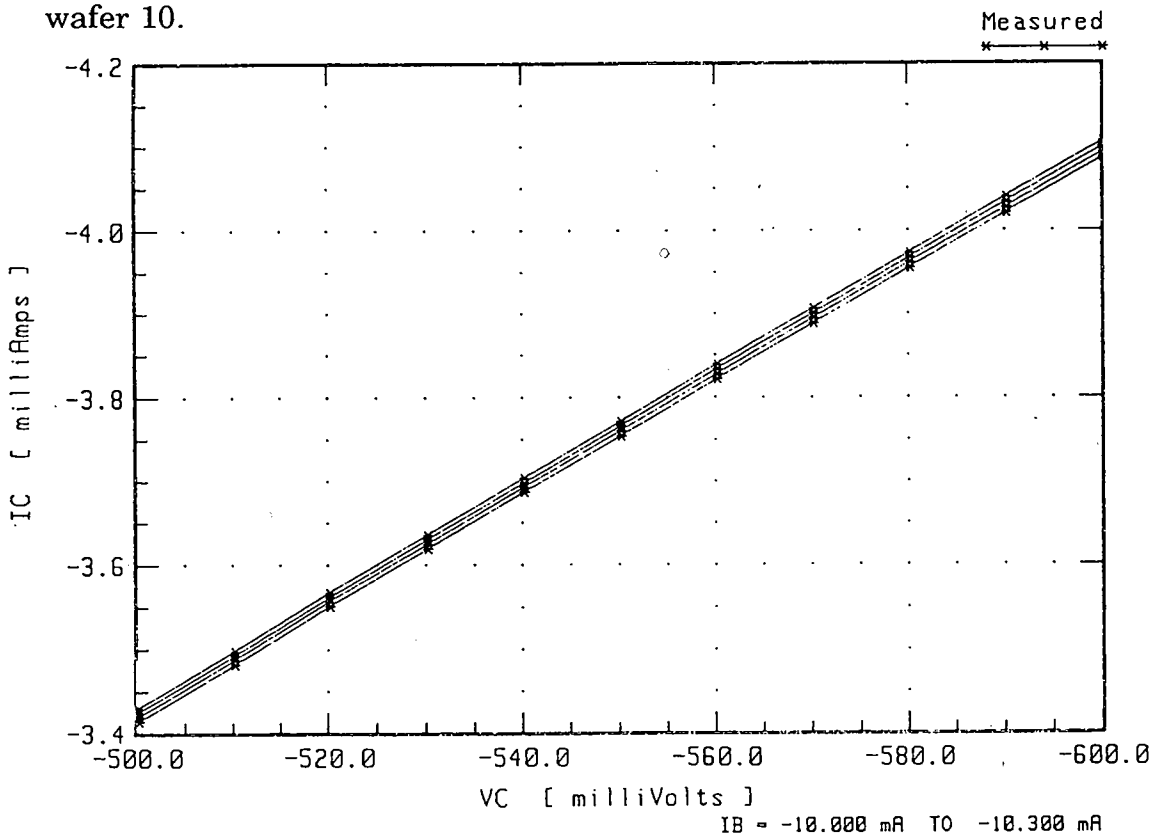


Figure 7.2h Typical measured curves for R_C extraction, taken from wafer 10.

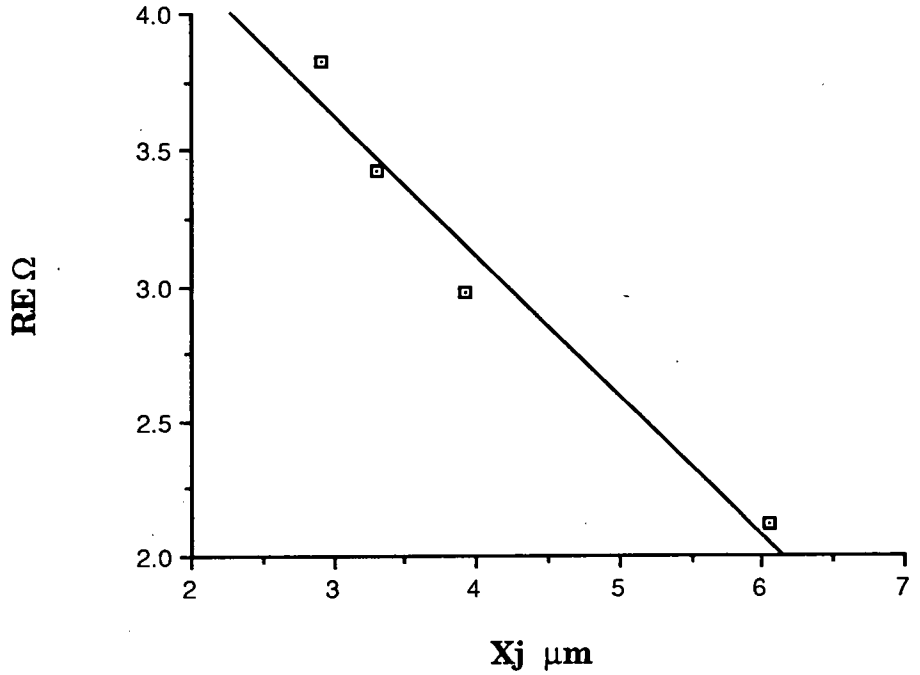


FIGURE 7.3a Plot of mean emitter resistance vs n-well junction depth.

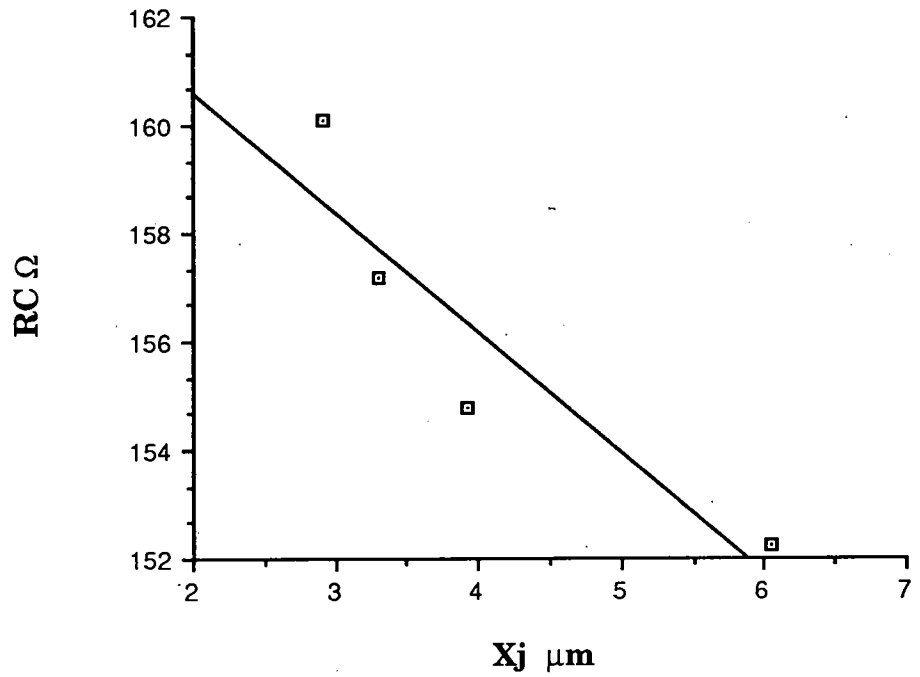


FIGURE 7.3b Plot of mean collector resistance vs n-well junction depth.

space charge region at the base-collector junction will increase in width as the base of the device is driven in longer. This is due to the decreased spatial distribution of dopant in the base, at the junction. All devices were characterised using a collector voltage sweep from -500mV to -600mV and stepped base current from -10mA to -10.3mA. The base current is high enough to ensure that the device is saturated but small enough to ensure that the device is not damaged by the power dissipated. R_C will vary with I_C increasing from a small value (R_{Csat}) to a larger value ($R_{Cnormal}$) as the collector current saturates. The observed variation in R_C is due to the devices with the shallower base reaching collector current saturation, and consequently $R_{Cnormal}$, before those with the deeper base junction. If a different region of extraction had been chosen for each drive-in split, a constant collector resistance could have been found for the devices.

7.1.5 Early Voltage

Table 7.3 shows the SMU set-up for extracting the forward and reverse Early voltages from the devices. Figures 7.4 a-d show typical measured curves for each part of the drive-in split. Figure 7.5a-b shows the wafer mean of V_{AF} and V_{AR} plotted against base junction depth. Figure 7.5a shows an increase of V_{AF} with drive-in time. Figure 7.5b also shows an increase in V_{AR} with drive-in time. Chapter 2 discussed the Early effect and chapter 3 discussed the parameters V_{AF} and V_{AR} used to model the Early effect. By considering these discussions we can get a qualitative feel for the relationship between the Early voltages, V_{AF} and V_{AR}, and drive-in time. The Early effect occurs when the base-collector space charge region (in the case of V_{AF}) encroaches into the base region reducing the width of the active base. A smaller early voltage indicates a greater effect on base width. If we assume a constant encroachment for a particular value of V_{CB} (not strictly true as the base-collector junction profiles are different for each part of the split) then the devices with the narrowest base widths (shortest drive-in) will be affected much more by this encroachment. This is illustrated by figure 7.5a.

7.1.6 Forward Gummel Parameters

Table 7.4 details the SMU set up for the extraction of the forward

Setup Name : Early Voltage	Setup # 3		
Function	MAIN	STEP	CONSTANT
Source Name	VC	IB	VE
Sweep Mode	LIN	LIN	
Start	2V	-1μ A	0.0 V
Stop	-4V	-10μ A	
# of Points	20	3	
Compliance	100 mA	3.0 V	100 mA
Fixed Sources			
Outputs	IC		

Table 7.3 SMU set up for the measurement of V_{AF} and V_{AR}.

Setup Name : IcIb vs Vb	Setup # 4		
Function	MAIN	CONSTANT	CONSTANT
Source Name	VB	VC	VE
Sweep Mode	LIN		
Start	-380 mV	-5 V	0.0 V
Stop	-900 mV		
# of Points	22		
Compliance	100 mA	100 mA	100 mA
Fixed Sources			
Outputs	IB	IC	

Table 7.4 SMU set up for the measurement of the forward Gummel parameters.

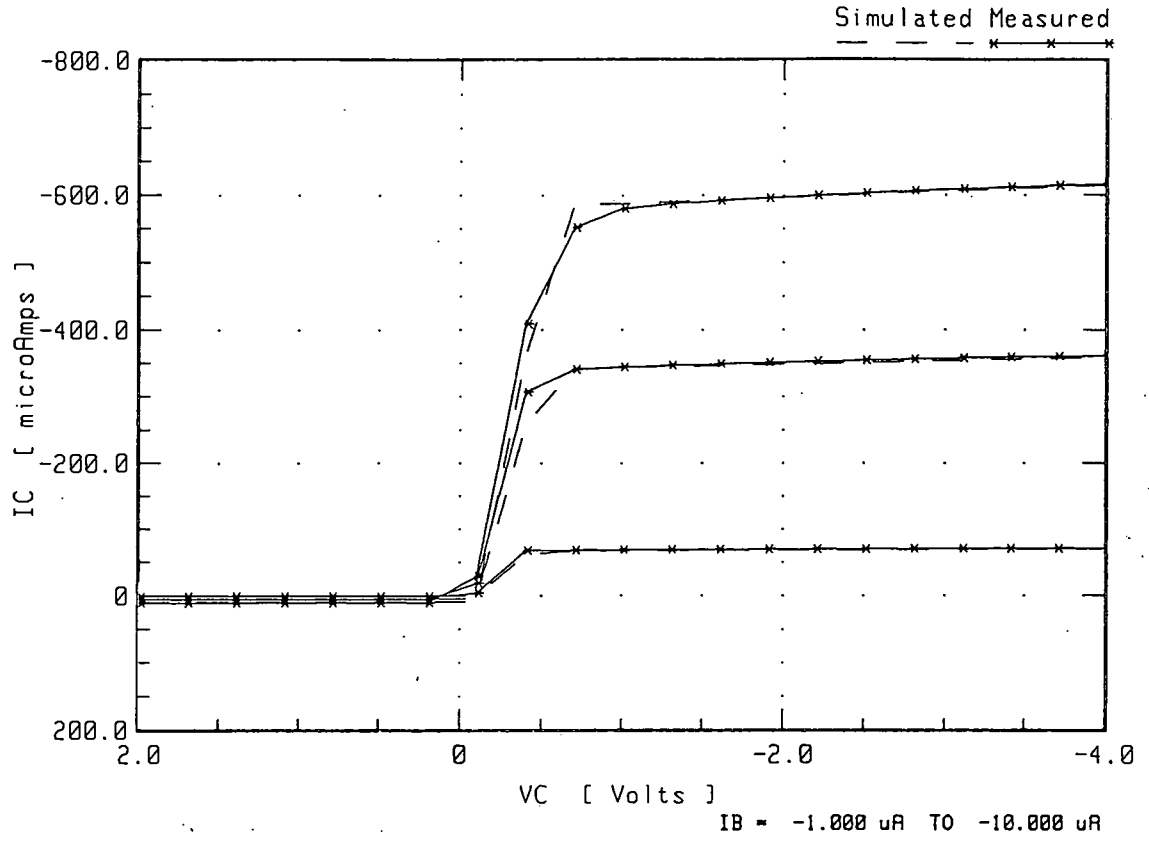


Figure 7.4a Typical measured curves for V_{AF} and V_{AR} extraction, taken from wafer 1.

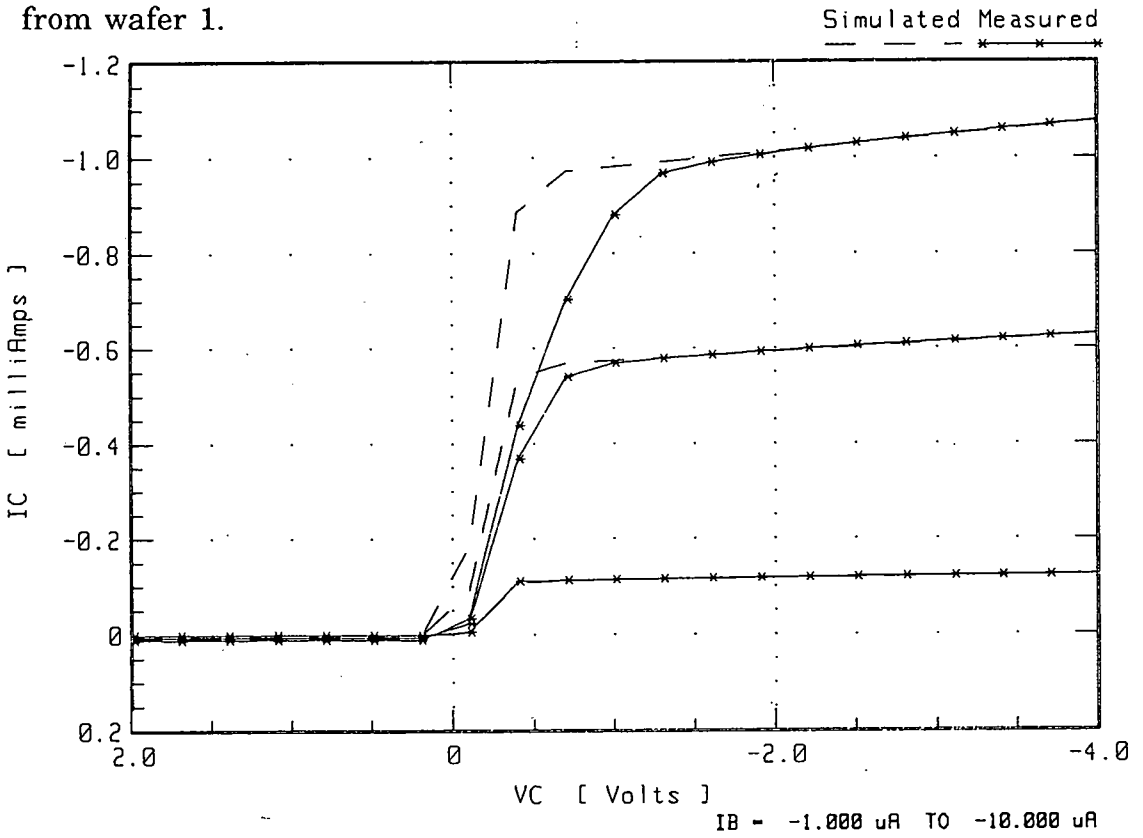


Figure 7.4b Typical measured curves for V_{AF} and V_{AR} extraction, taken from wafer 4.

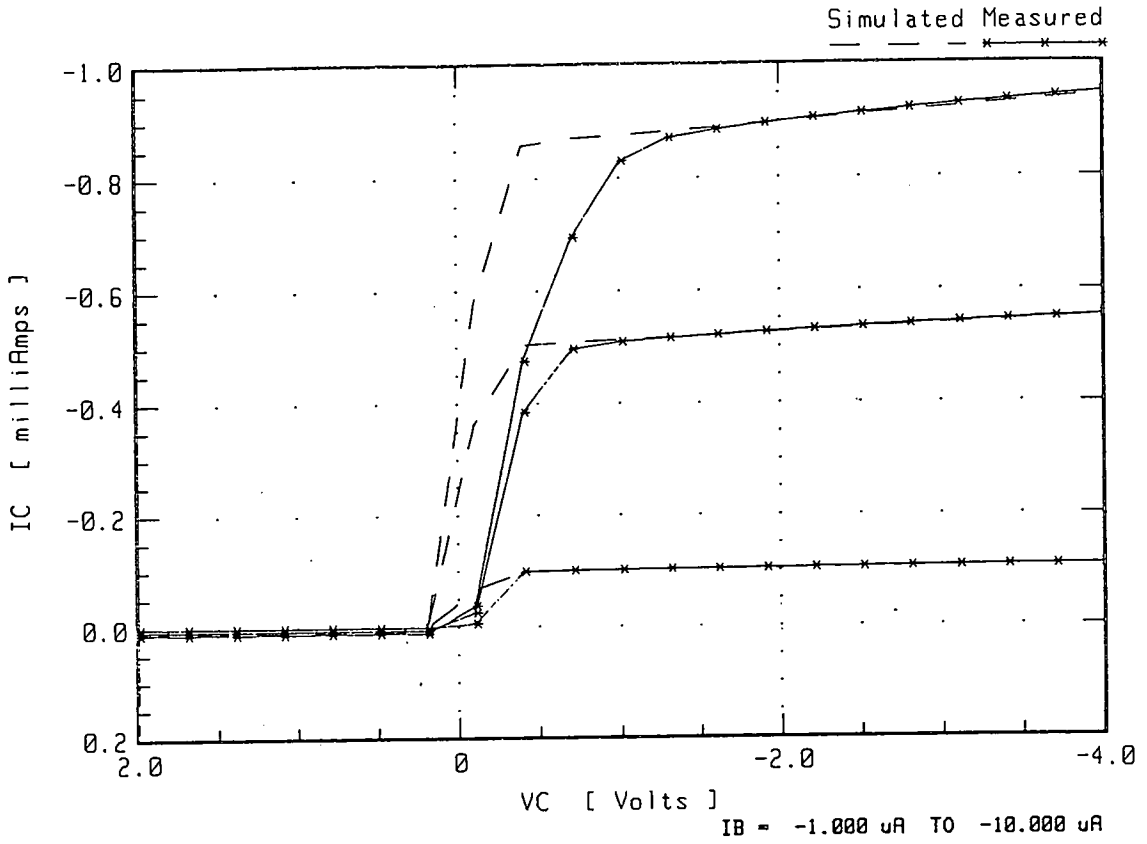


Figure 7.4c Typical measured curves for V_{AF} and V_{AR} extraction, taken from wafer 6.

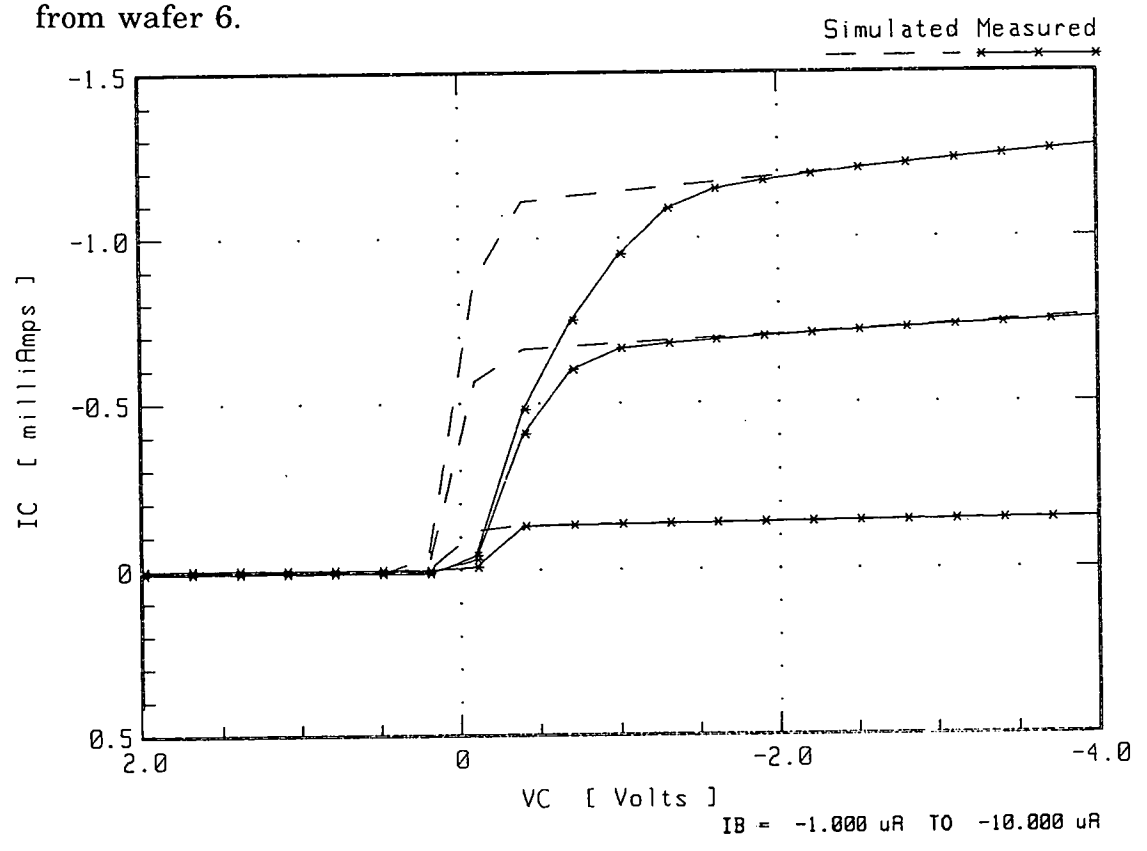


Figure 7.4d Typical measured curves for V_{AF} and V_{AR} extraction, taken from wafer 10.

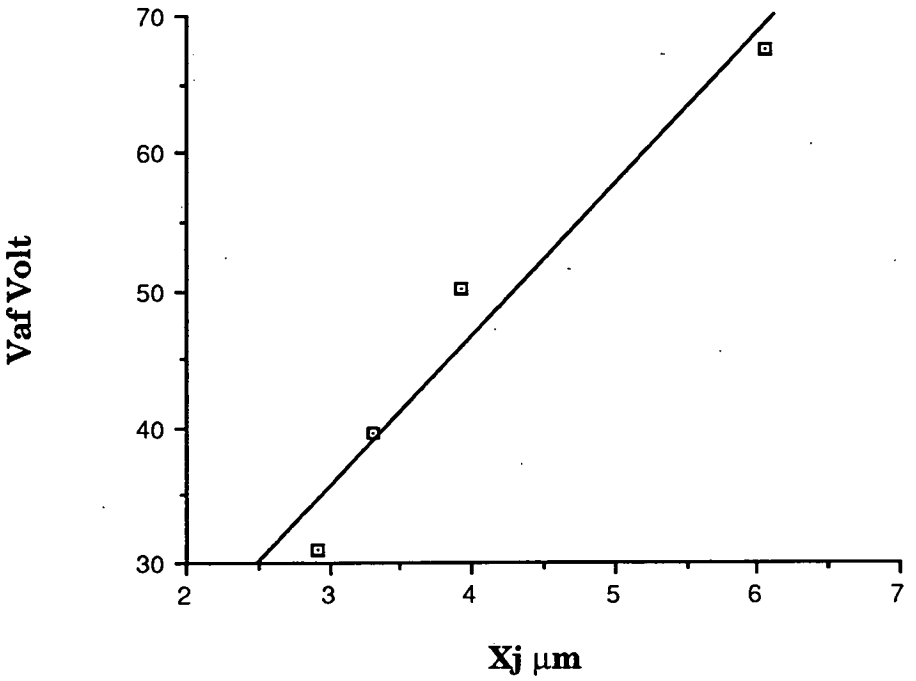


Figure 7.5a Plot of forward Early voltage vs n-well junction depth.

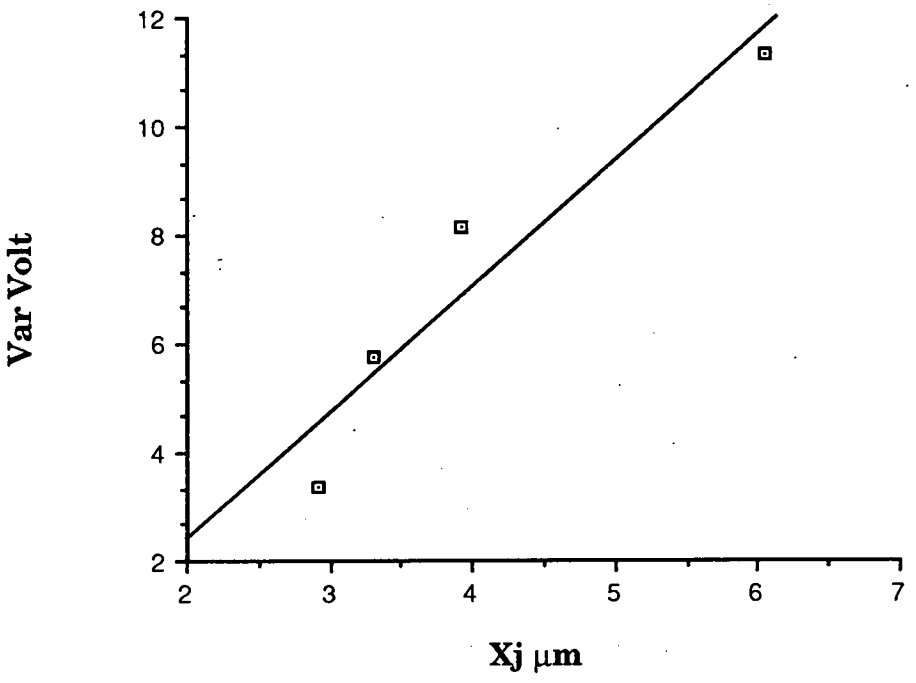


Figure 7.5b Plot of mean reverse Early voltage vs n-well junction depth.

Gummel parameters I_S, N_F, β_F, I_{KF}, I_{SE}, N_E, and R_B. A detailed derivation of these parameters is given in chapter 3. Figure 7.6 a-d shows the typical extracted forward Gummel curves for each part of the drive-in split. Figures 7.7a-d show typical forward biased gain characteristics for each drive-in condition. Figures 7.8a-g show the wafer mean for each extracted parameter vs base junction depth.

Figure 7.8a shows the fall of I_S (transport saturation current) with increased base junction depth. As the base junction is driven in, the built-in base charge Q_{BO} will increase. From equation 3.21, as Q_{BO} rises then I_S will fall. This is a consequence of the rise in the Gummel Number.

Figure 7.8b shows the fall of N_F (forward current emission coefficient) with increased base junction depth. N_F models the deviation of the emitter-base diode from ideal. The closer the parameter is to unity the more ideal the junction. As the base junction is driven in the effective base doping at the emitter-base junction will fall. The ratio of emitter-base doping will increase. The recombination in the emitter-base space charge region will decrease and thus the junction will become more ideal.

Figure 7.8c shows the fall of β_F (forward current gain) with increased base junction depth. From the theory presented in chapter 2 this is a fairly obvious relationship. The gain is inversely proportional to the active base width. Another effect touched on above is that of the increased Gummel Number, this also serves to reduce the gain of the transistor. However, this is a much smaller order effect when compared with base width variation.

Figure 7.8d shows the relationship of I_{KF} (knee current for forward beta high current roll off) vs base width. As the base width increases I_{KF} falls. This is a consequence of reduced gain of the devices with the larger base widths. As they begin to run into high current effects at the emitter-base junction, less of the injected current reaches the collector.

Figures 7.8e and 7.8f show the relationship of I_{SE} (base-emitter leakage saturation current) and N_E (base-emitter leakage emission coefficient) vs base width. Both of these parameters fall with increased base width. The fall of I_{SE} is a consequence of decreased recombination in the emitter-base space charge region. There are other recombination

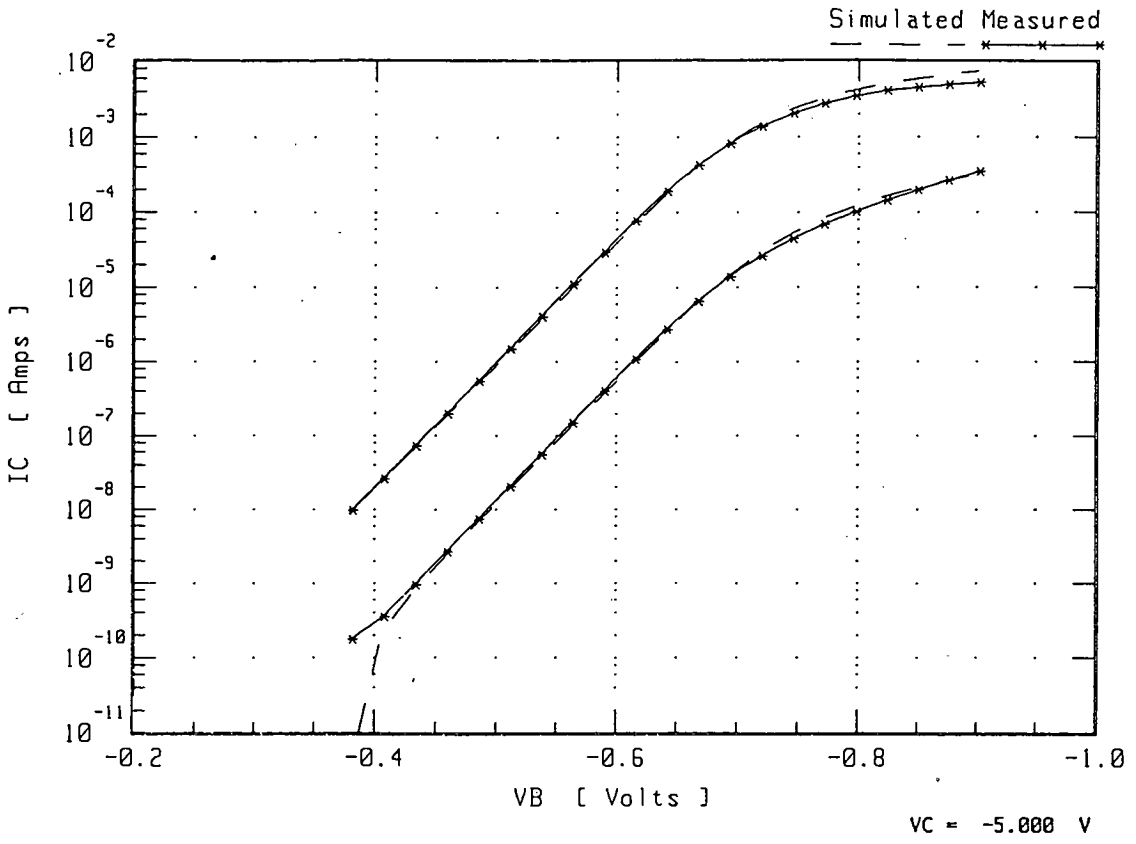


Figure 7.6a Typical measured forward Gummel curves for I_S, N_F, β_F, I_{KF}, I_{SE}, N_E, and R_B. extraction, taken from wafer 1.

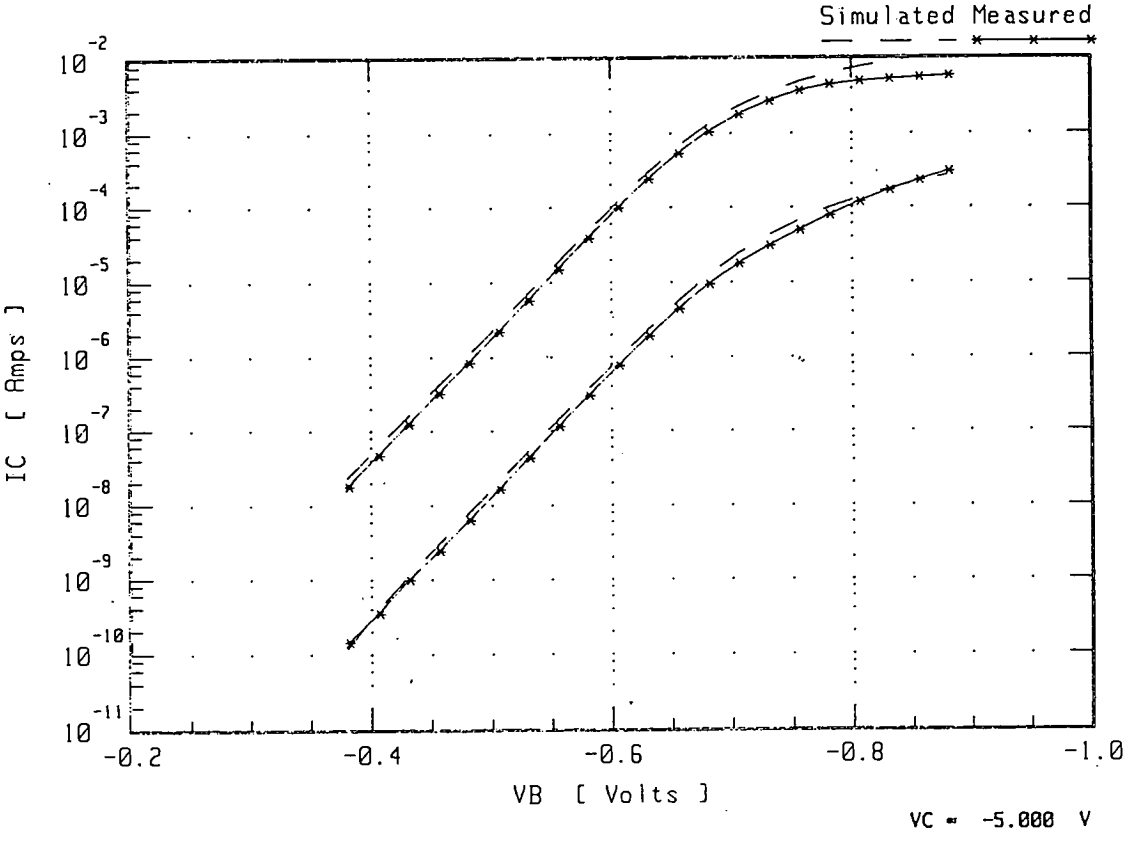


Figure 7.6b Typical measured forward Gummel curves for I_S, N_F, β_F, I_{KF}, I_{SE}, N_E, and R_B. extraction, taken from wafer 4.

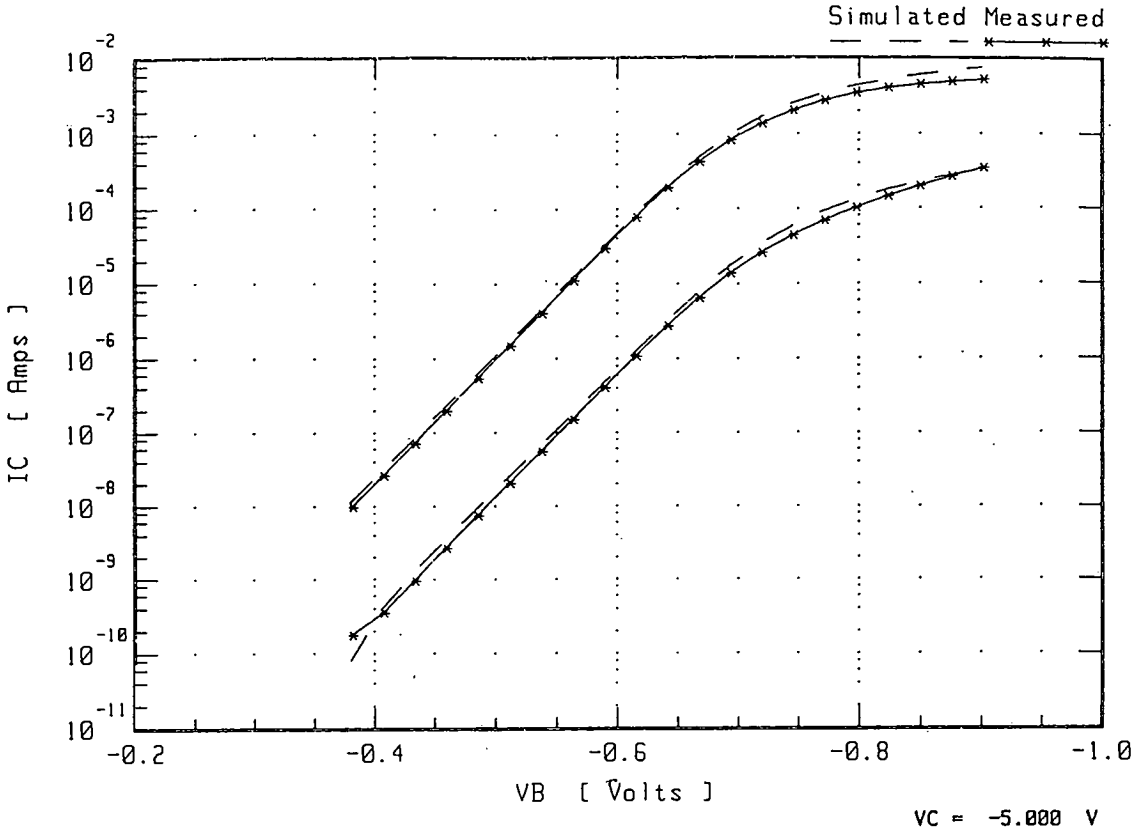


Figure 7.6c Typical measured forward Gummel curves for I_S, N_F, β_F, I_{KF}, I_{SE}, N_E, and R_B. extraction, taken from wafer 6.

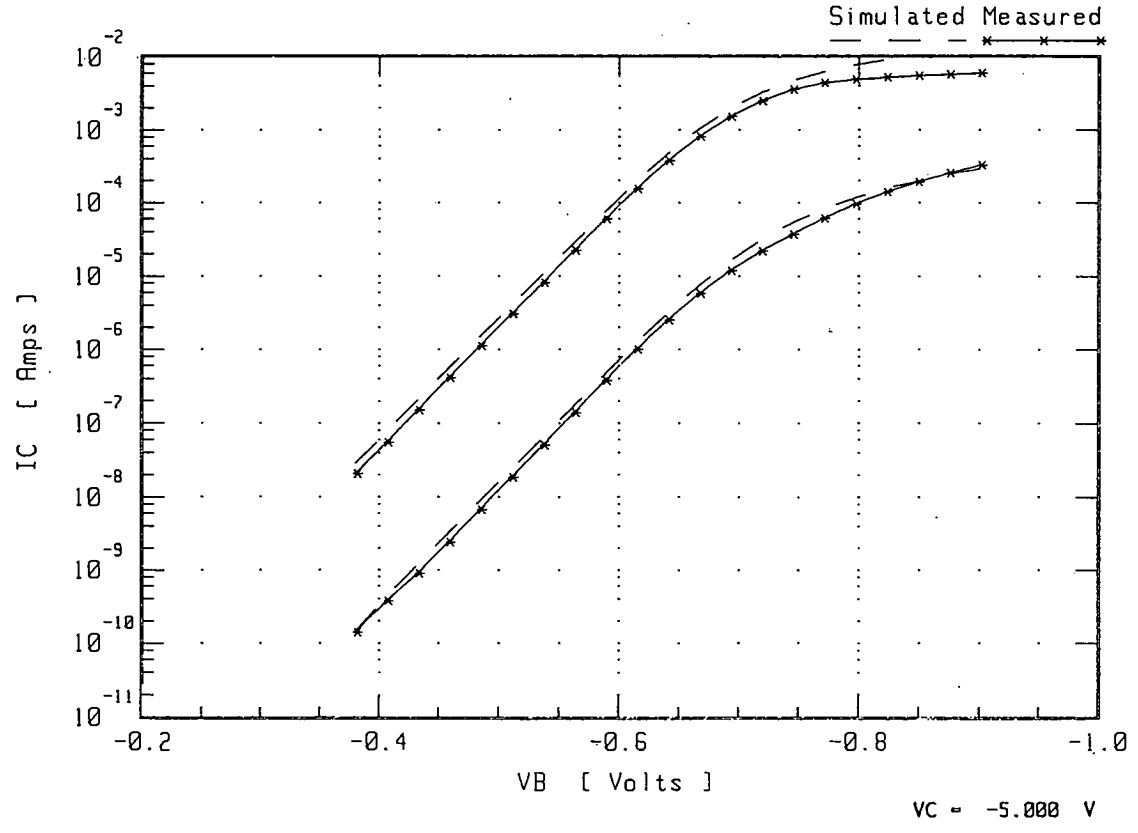


Figure 7.6d Typical measured forward Gummel curves for I_S, N_F, β_F, I_{KF}, I_{SE}, N_E, and R_B. extraction, taken from wafer 10.

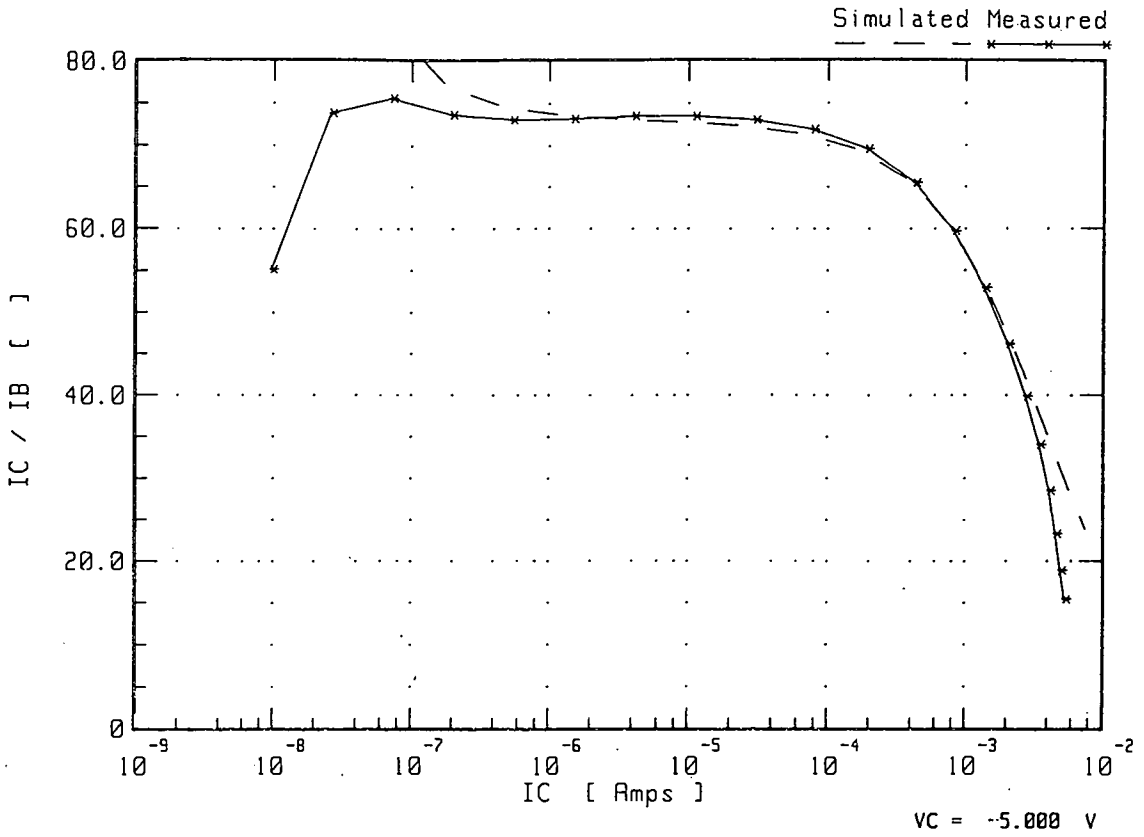


Figure 7.7a Typical measured forward β curve taken from wafer 1.

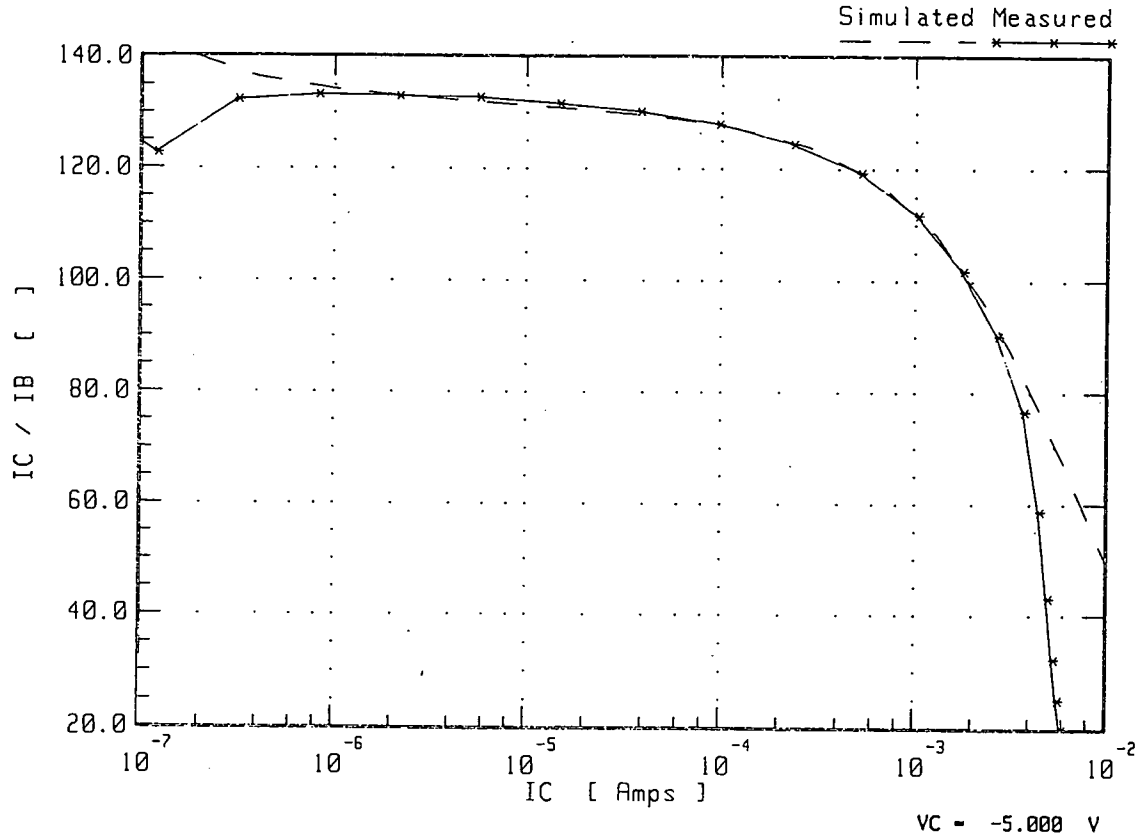


Figure 7.7b Typical measured forward β curve taken from wafer 4.

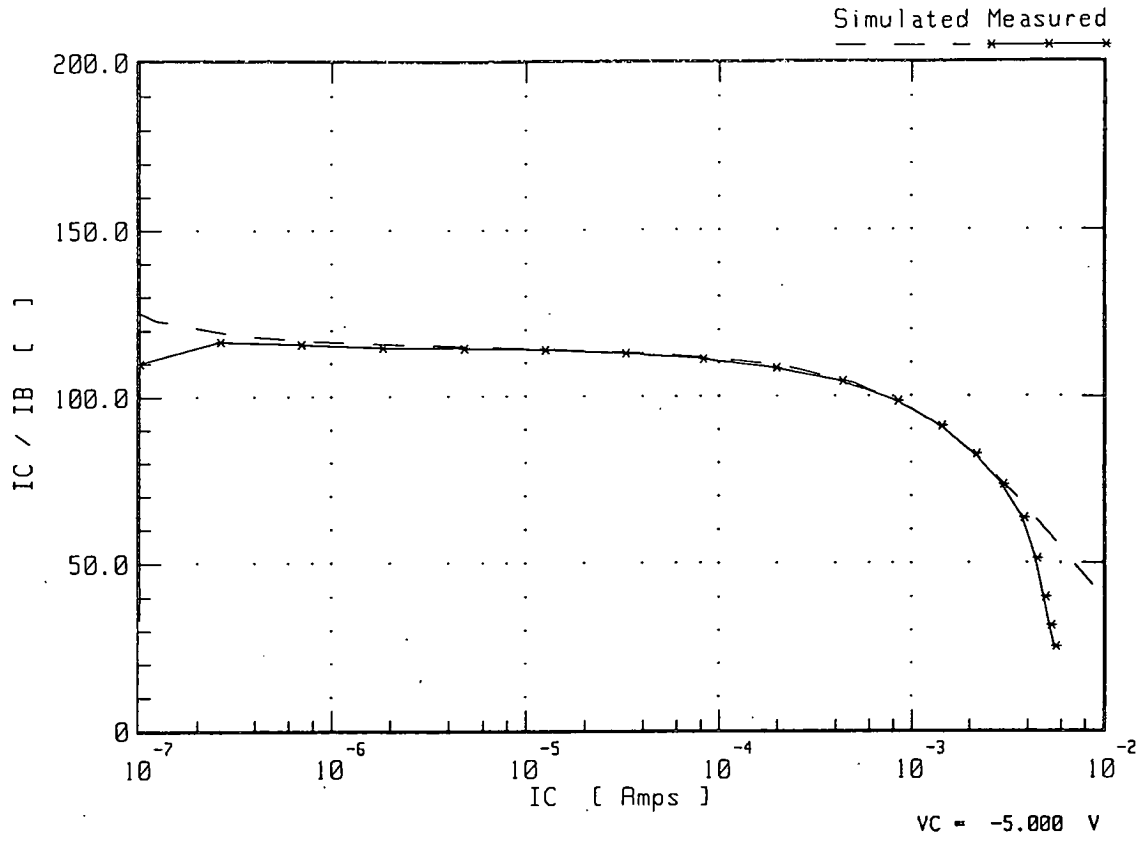


Figure 7.7c Typical measured forward β curve taken from wafer 6.

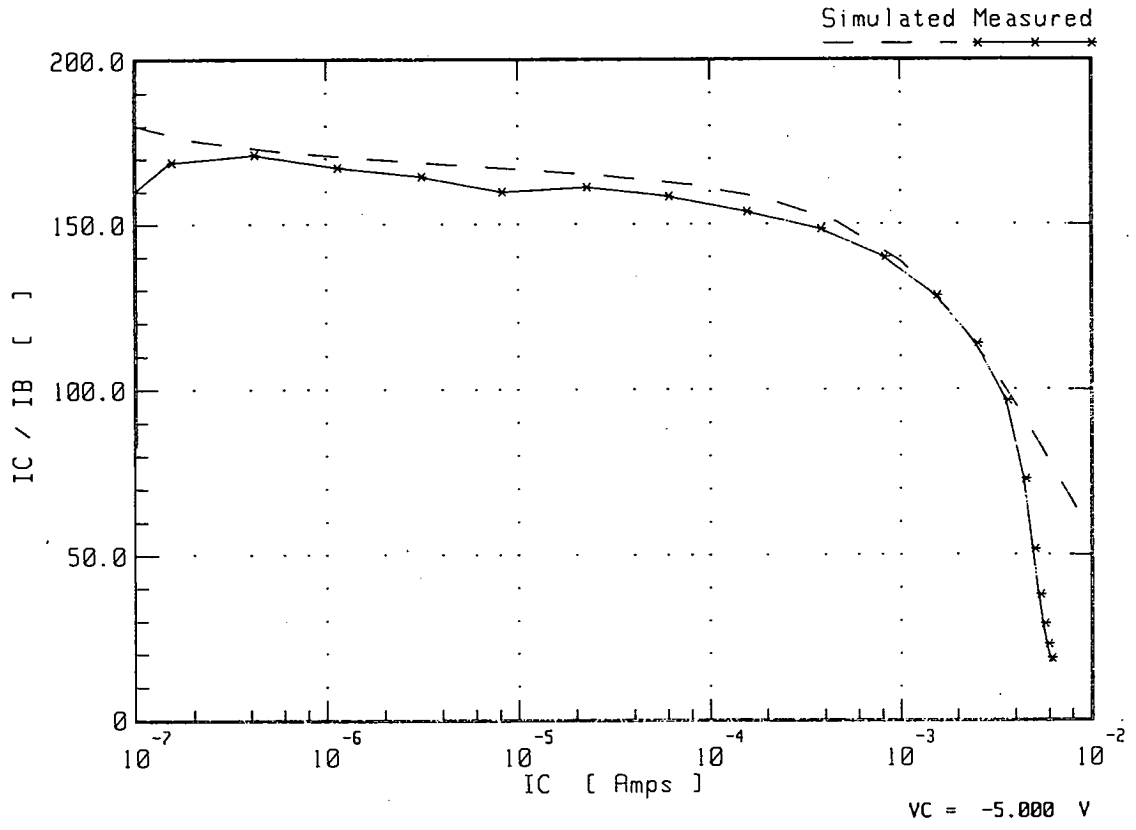


Figure 7.7d Typical measured forward β curve taken from wafer 10.

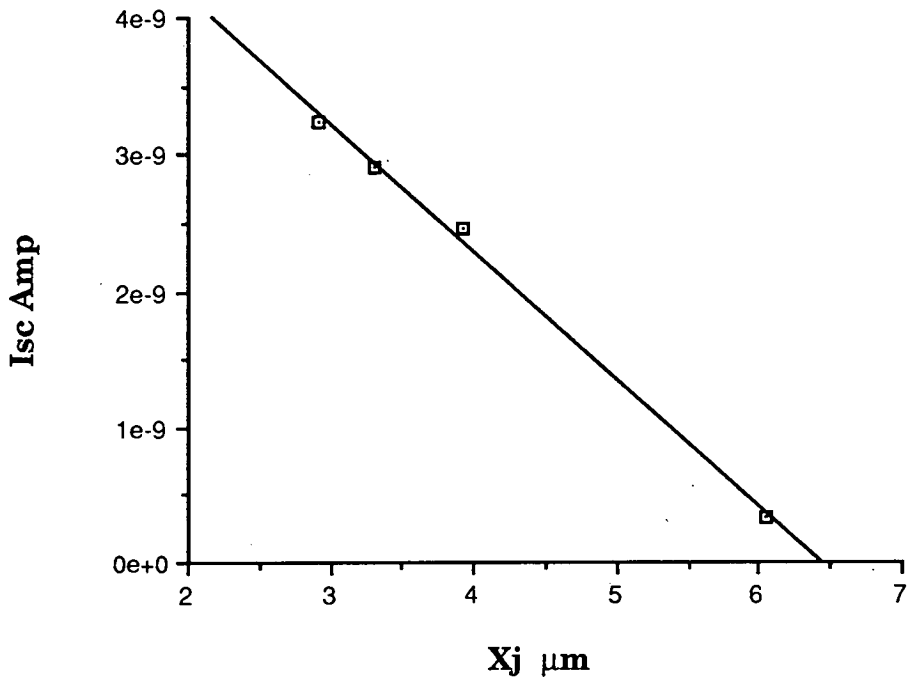


Figure 7.8a. Plot of I_S (transport saturation current) vs n-well junction depth.

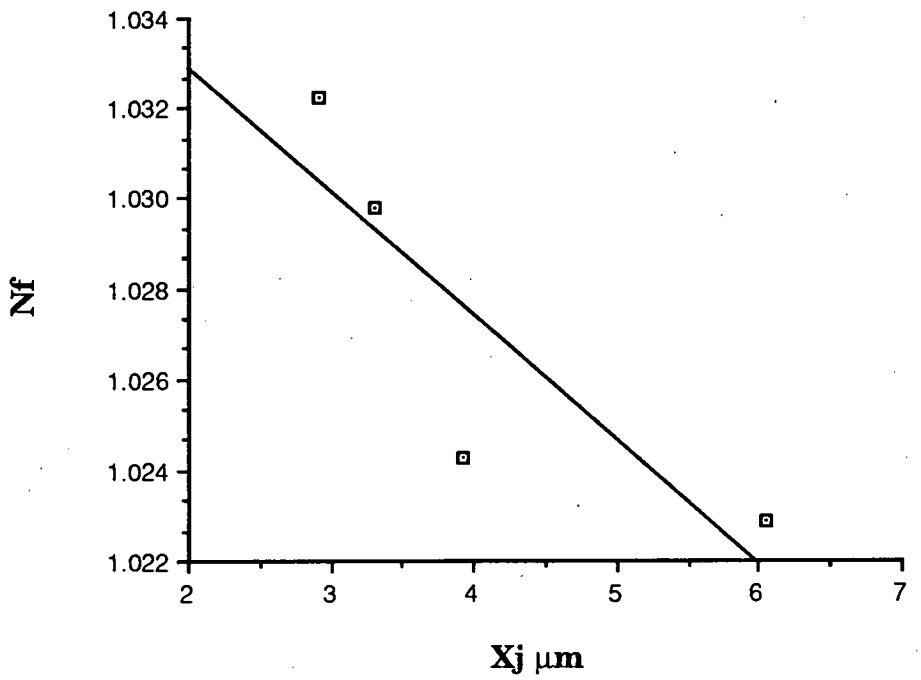


Figure 7.8b Plot of N_F (forward current emission coefficient) vs n-well junction depth.

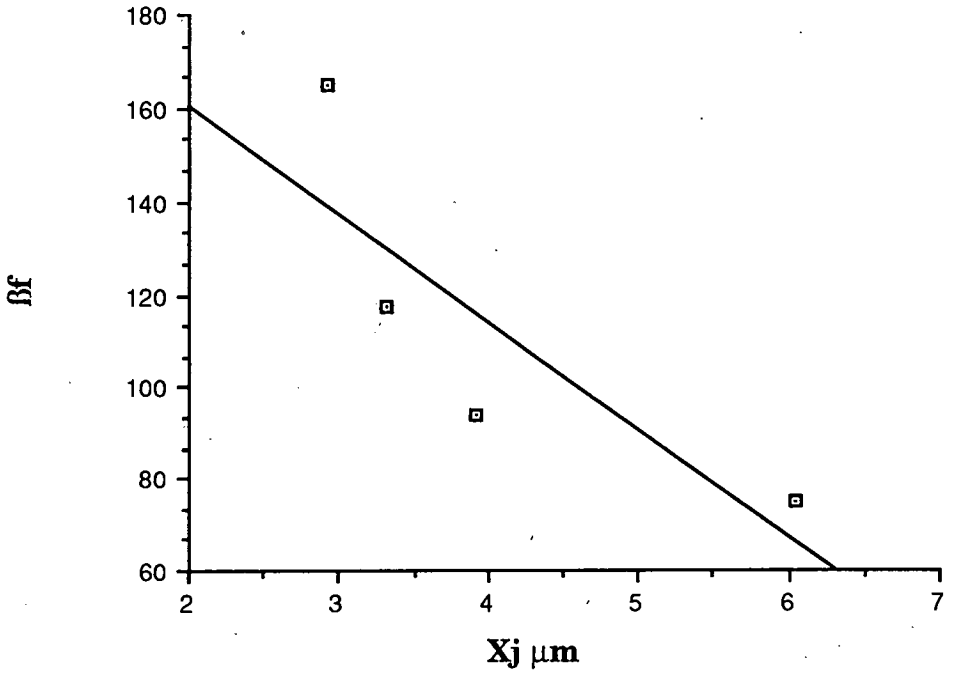


Figure 7.8c Plot of β_F(Ideal maximum forward beta) vs n-well junction depth.

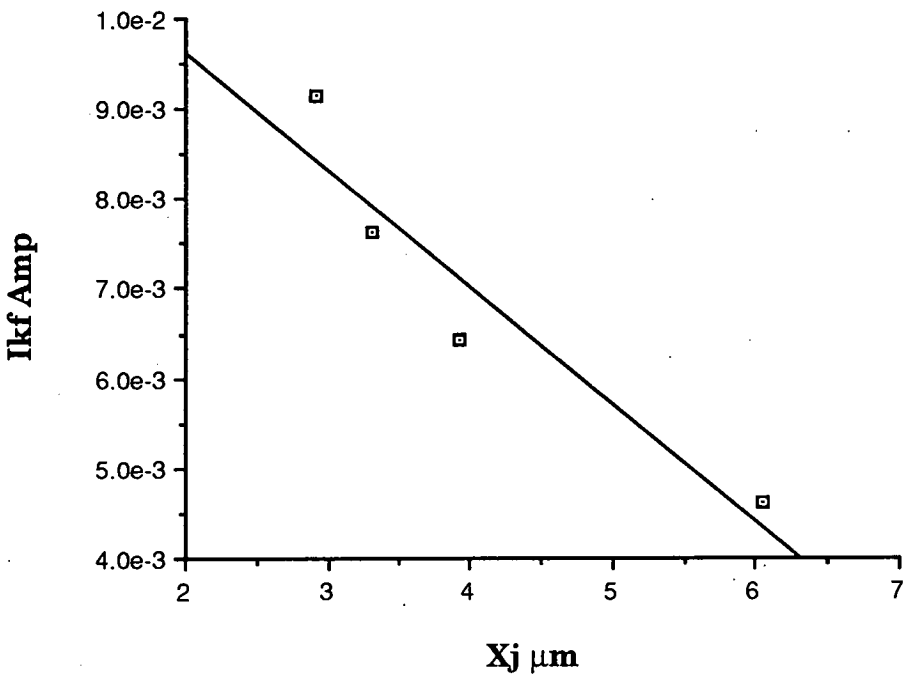


Figure 7.8d Plot of I_{KF} (Knee current for forward beta roll off) vs n-well junction depth.

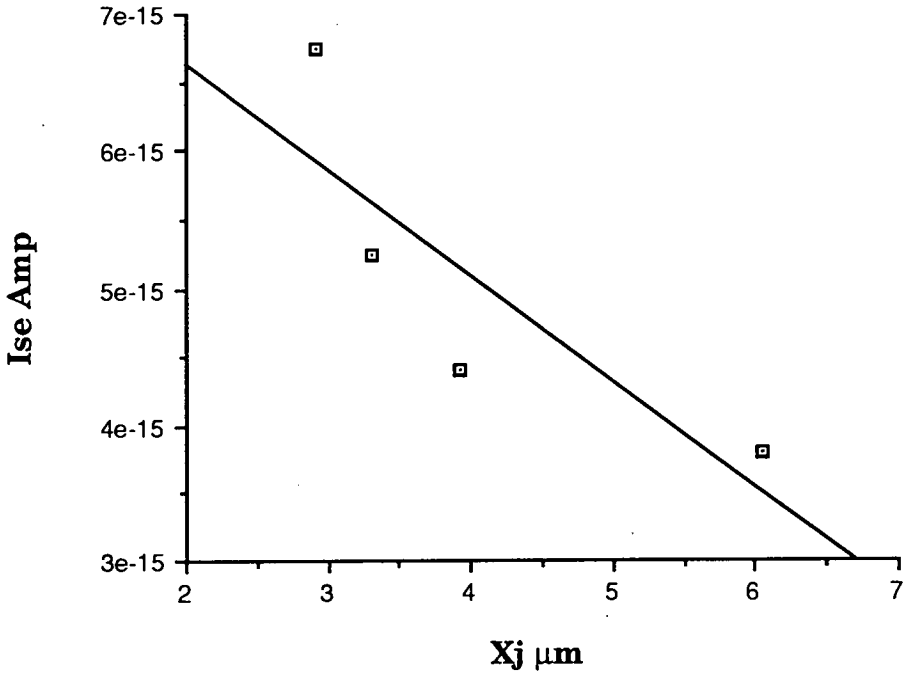


Figure 7.8e Plot of I_{SE} (base-emitter leakage saturation current) vs n-well junction depth.

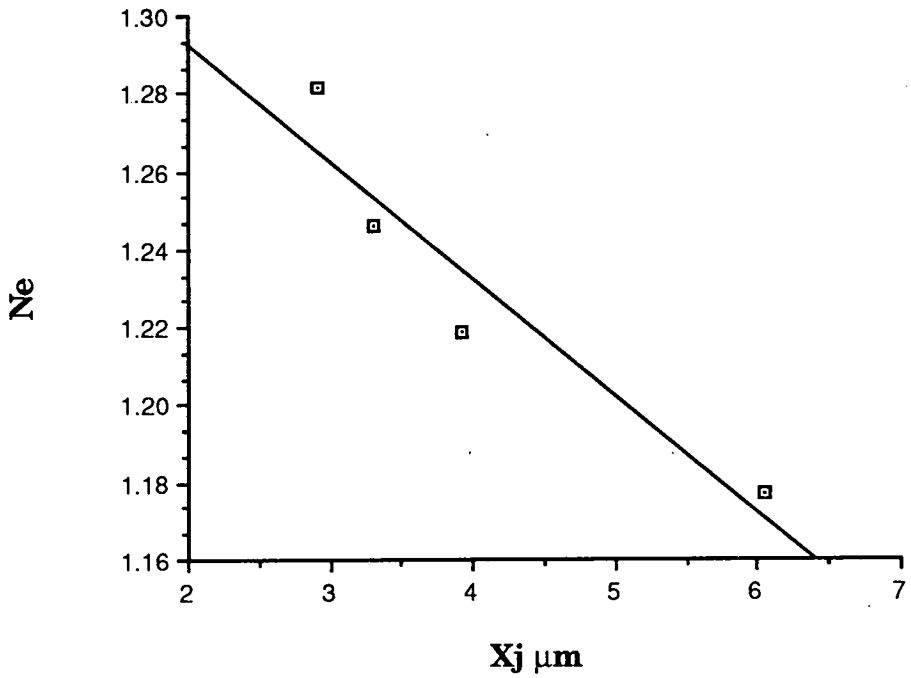


Figure 7.8f Plot of N_E (base-emitter leakage emission coefficient) vs n-well junction depth.

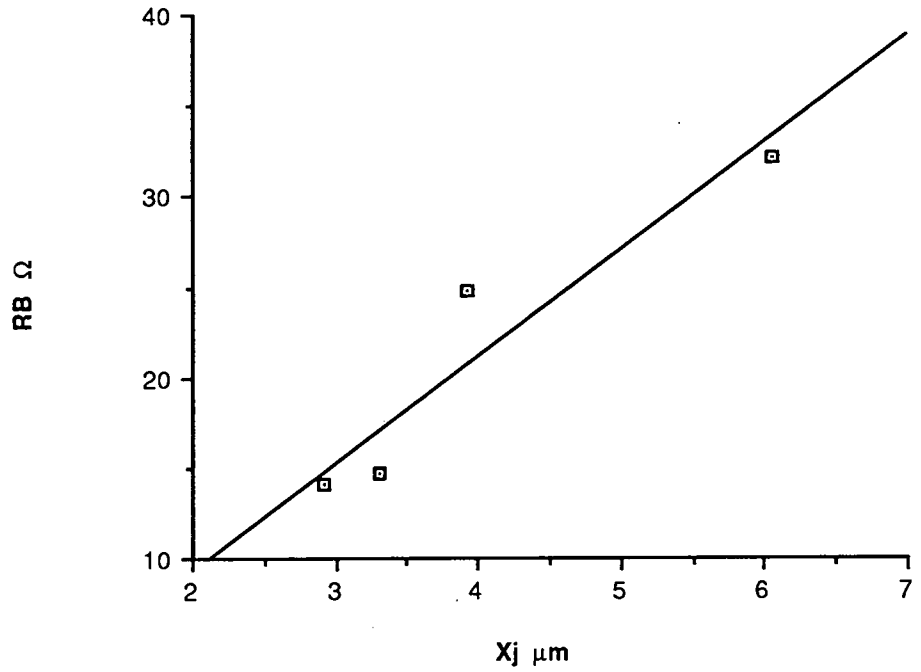


Figure 7.8g Plot of R_B (zero bias base resistance) vs n-well junction depth

effects at the surface but the recombination in the space-charge region dominates. These effects are only significant at low currents.

Figure 7.8g shows the relationship of R_B (zero bias base resistance) vs base junction depth. As the base junction depth increases, the parasitic resistance in the base increases. Although, as noted above, the active base charge increases with drive-in time this charge is distributed spatially over a much greater area. This causes the observed increase in parasitic base resistance.

7.1.7 Reverse Gummel Parameters

Table 7.5 details the SMU set-up for the extraction of the reverse Gummel parameters (N_R, β_R, I_{KR}, I_{SC}). Figure 7.9a-d shows the wafer mean of each extracted parameter vs base junction depth.

Figure 7.9a shows the relationship of N_R (reverse current emission coefficient) vs base junction depth. As detailed in chapter 3, N_R models the deviation of the base-collector diode from the ideal. The closer the parameter is to unity the more ideal the junction. Typically in

discrete or IC bipolar devices this is very close to unity. As the base junction is driven in, N_R falls from around 2 to about 1.75. From this we can say that the base-collector diode is never very ideal. However, as the base junction is driven in, doping concentration at the base side of the junction falls. The situation becomes like that of N_F dealt with above.

Figure 7.9b shows β_R (maximum reverse gain) vs base junction depth. β_R falls as the base junction is driven in. This follows the fall of β_F with base junction depth. Here the critical factor is base width. As the base width increases the gain of the device falls.

Figure 7.9c shows I_{KR} (knee current for reverse beta high current roll off) vs base junction depth.Figure 7.11d illustrates I_{SC} (base-collector leakage saturation current) vs base junction depth. As the base is driven in we have seen, from N_R above, that the base-collector diode becomes more ideal. As with I_{SE} above the recombination in the base-collector space charge region falls. This results in a decrease in I_{SC}.

The above parameter relationships illustrate that the parasitic bipolar transistor is sensitive to variations in the CMOS process. These parameters vary in a predictable and understandable manner. It will be shown later that this sensitivity can be utilised to evaluate and therefore control CMOS process uniformity.

7.2 The MOS Transistor

MOS transistor arrays were included in the design of the test chip (chapter 4). These were included to allow comparisons of device sensitivity to process change and also to explore the relationship between extracted MOS parameters and those of the parasitic bipolars. The results presented below are extracted from devices on the same wafers as those shown above. Typical characteristics are shown for one MOS characterisation and are presented with wafer averages of the parameters extracted for each part of the split.

7.2.1 Device Structure

Although both p- and n-channel devices were designed and fabricated only p-channel results are presented below. The reasons for this are twofold.

1. The vertical bipolar parasitics and JFET devices with which the MOS transistors will be compared were fabricated using the n-well. Any comparisons between these devices and n-channel MOSFETs would be of very little use.

2. Although the EMF 5μm CMOS process is a single well process, any relationships developed here would be applicable to a twin well technology where parasitic vertical pnp and npn parasitics are available as test structures.

7.2.2 Characterisation Technique

Parameters for the SPICE level 2 MOS model were extracted from the MOS transistor arrays. However, using TECAP, parameters for this model are achieved by running the optimizer (chapter 6) to provide a best fit of the model to the extracted data, in the appropriate region of operation. This had to be borne in mind when making any comparisons with other devices.

The first step in extracting parameters for the MOS model was to enter values for T_{OX}, L_D, W_D and R_S. T_{OX} is the gate oxide thickness for the process, this is used in the calculation of the conduction factor and back-gate bias effects. L_D is the lateral diffusion coefficient (refered to as ΔL in chapter 6). This is used to determine the initial value of the effective channel length. W_D is the channel width reduction factor and this is used to determine the value of the effective channel width. This parameter exists only in TECAP and is assumed to be zero in the SPICE model. R_S is the source/drain ohmic resistance. This parameter is geometry independent in both SPICE and TECAP. X_j is the metallurgical source/drain junction depth, that is the distance into the diffused region at which point the source/drain dopant concentration becomes equal to the background dopant concentration.

The next stage in the characterisation of the device is to extract the linear region parameters from a large device. From these curves the parameters μ_o, V_{TO}, N_{SUB}, Theta and N_{FS} are extracted. μ_o is the surface mobility at low gate voltages. V_{TO} is the extrapolated zero bias threshold voltage (chapters 2 and 3).

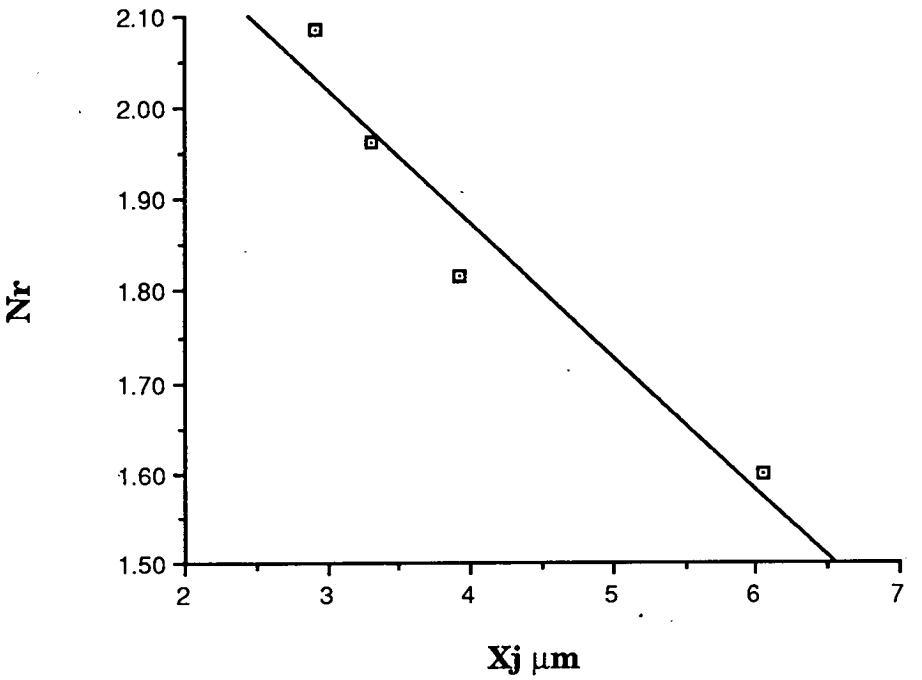


Figure 7.9a. Plot of N_R (reverse current emission coefficient) vs n-well depth.

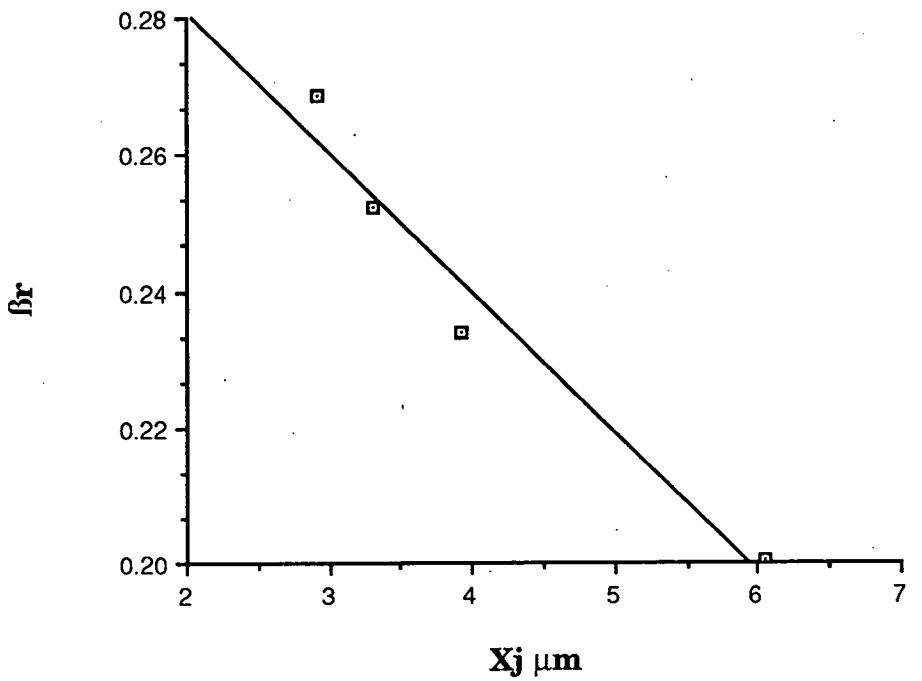


Figure 7.9b. Plot of β_R (Ideal maximum reverse beta) vs n-well depth.

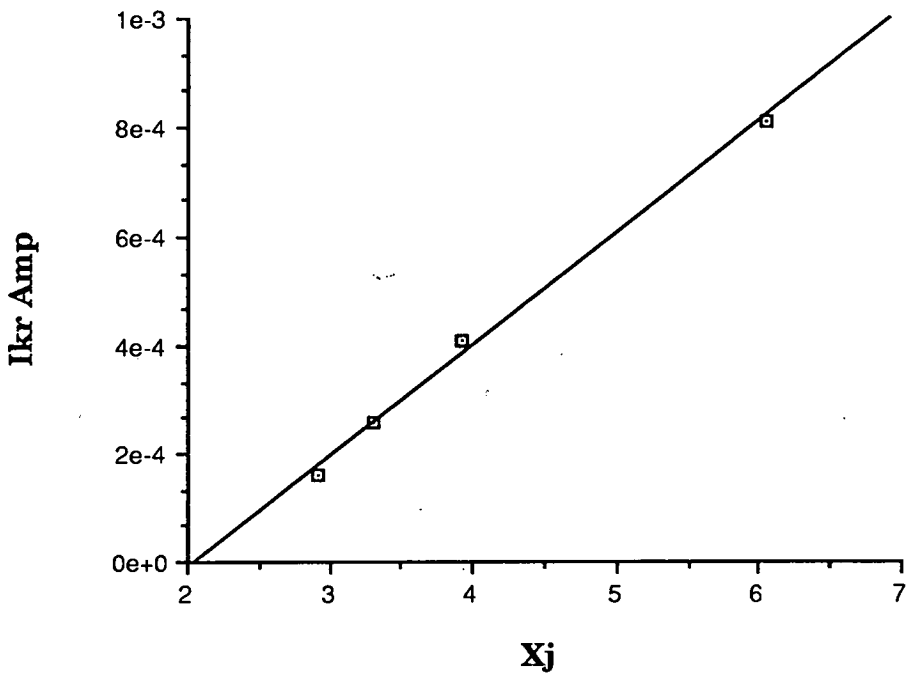


Figure 7.9c. Plot of I_{KR} (knee current for reverse beta high current roll off) vs n-well depth.

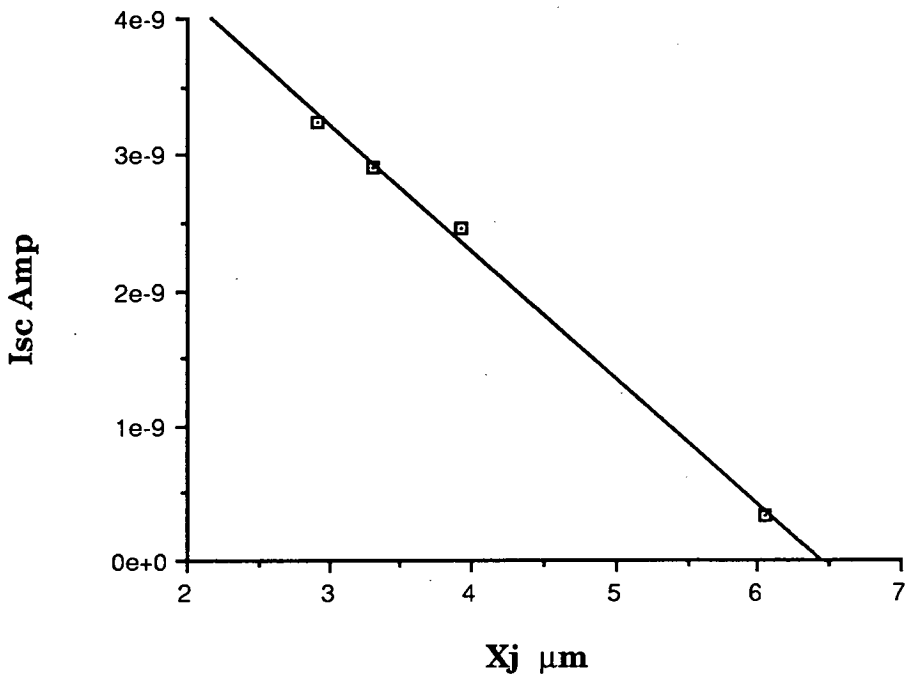


Figure 7.9d. Plot of I_{sc} (base-collector leakage saturation current) vs n-well depth.

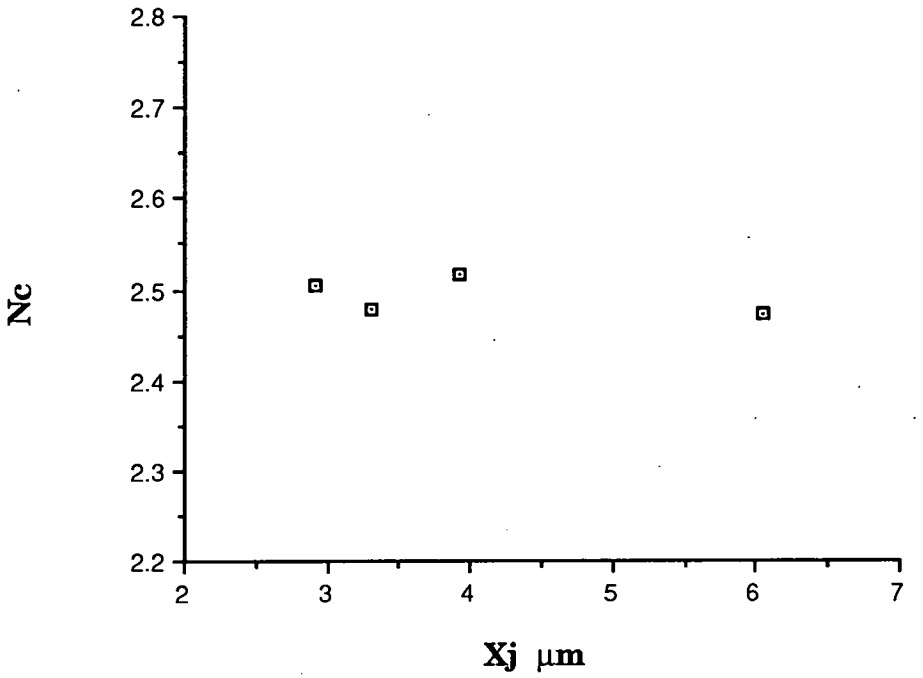


Figure 7.9e. Plot of N_C (base-collector leakage emission coefficient) vs n-well depth,

Setup Name : IeIb vs Vb	Setup #5		
Function	MAIN	CONSTANT	CONSTANT
Source Name	VB	VE	VC
Sweep Mode	LIN		
Start	-300 mV	-1.0 V	0.0 V
Stop	-900 mV		
# of Points	22		
Compliance	100 mA	100 mA	100 mA
Fixed Sources			
Outputs	IE	IB	

Table 7.5 SMU set up for the measurement of the reverse Gummel parameters.

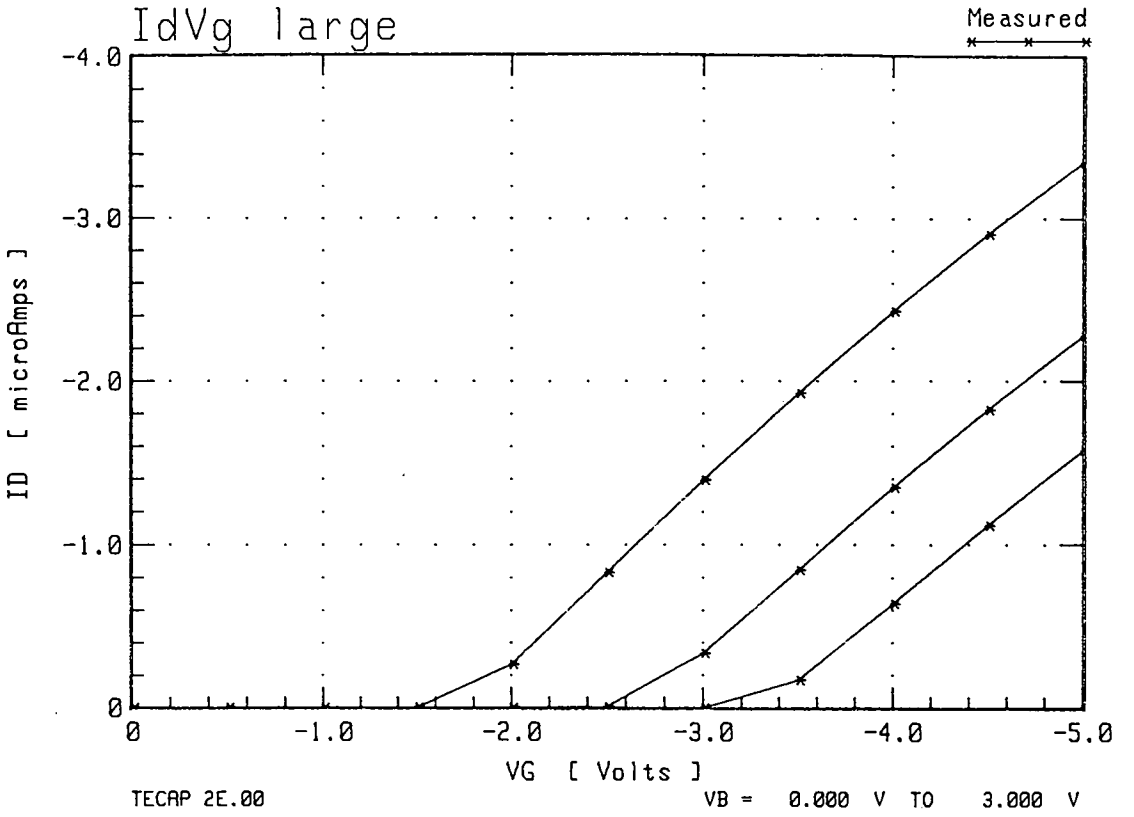


Figure 7.10a Typical characteristic curves for μ_o, V_{TO}, N_{SUB}, U_{EXP}, and N_{FS} extraction, taken from the 25µm x 25µm device on wafer 1.

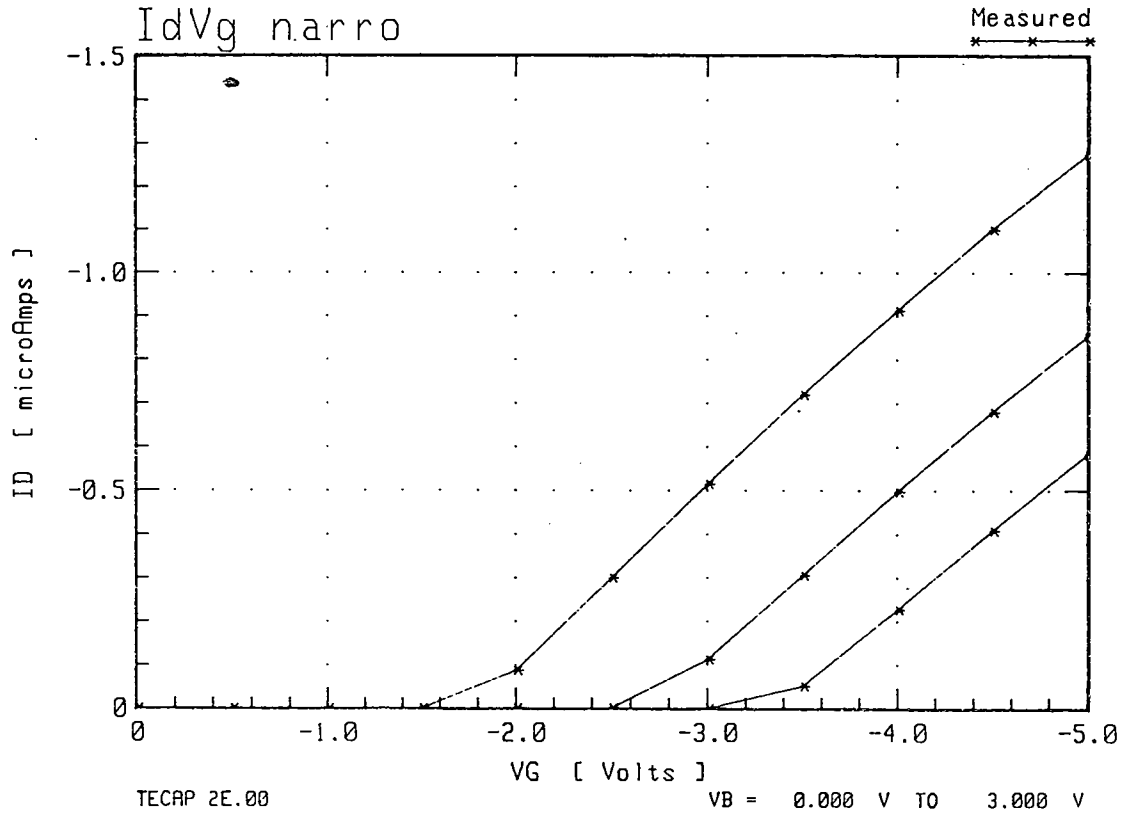


Figure 7.10b Typical characteristic curves for *WD* and DELTA extraction, taken from the 5µm x 25µm device on wafer 1.

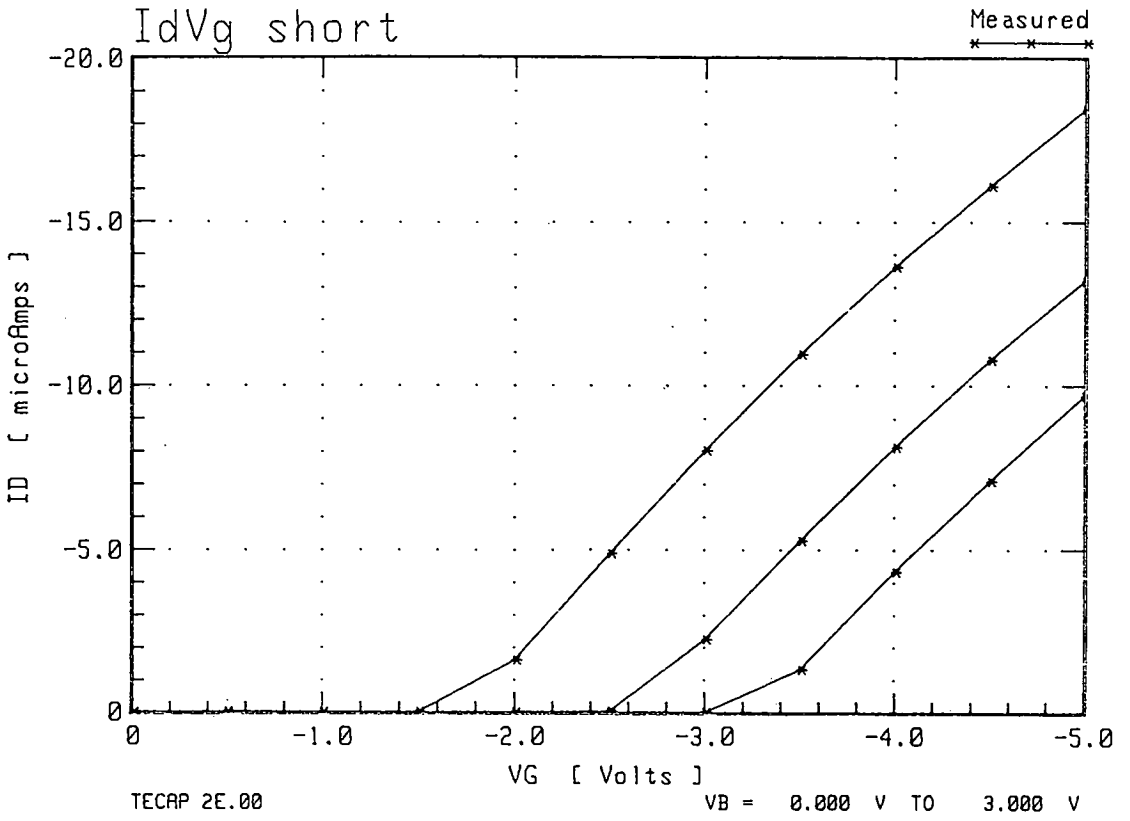


Figure 7.10c Typical characteristic curves for *LD, X_J, RD,* and*RS* extraction, taken from the 25µm x 5µm device on wafer 1.

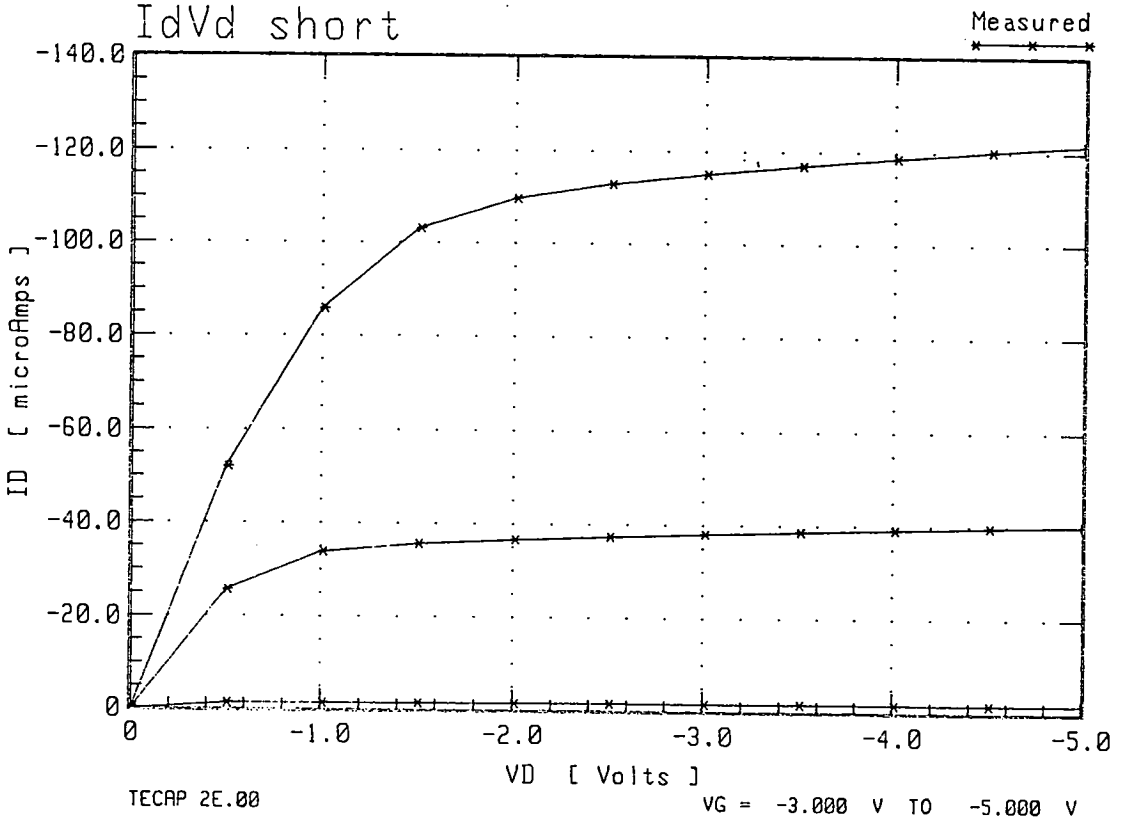


Figure 7.10d Typical characteristic curves for V_{MAX} and N_{EFF} extraction, taken from the 25µm x 5µm device on wafer 1.

This parameter models the onset of strong inversion and marks the point where the device starts conducting if weak inversion currents are neglected. N_{SUB} is the substrate doping concentration used in the calculation of most of the backgate bias effects. Theta is the critical field exponent and is used to model mobility reduction at high electrical field strengths in the channel. Theta is equivalent to U_{exp} discussed in chapter 3. N_{FS} is the effective surface state density and is used to determine sub-threshold current flow.

From a narrow device DELTA is extracted. This parameter models the effect of device width on threshold voltage.

The next stage in the characterisation is to extract V_{MAX} the maximum drift velocity of carriers. This determines whether V_{Dsat} is a function of scattering velocity limited carriers or the drain depletion region pinch-off. V_{Dsat} was discussed in chapter 3. Typical characteristic curves from a set of MOS devices on wafer 1 are given in figure 7.10a-d. In the following discussions no short or narrow channel parameters are used in comparison with the bipolar devices.

7.2.3 MOS Parameter Variation with Process Split

This section presents the variation of extracted MOS parameters with the process split discussed earlier in the chapter. The relationships shown are used in the later comparison with extracted bipolar parameters and a process uniformity evaluation. As above, the parameters presented are wafer mean parameters.

Figure 7.11a illustrates the variation of p-channel V_{TO} with simulated n-well junction depth. The graph shows a fall in $|V_{TO}|$ with n-well junction depth. This is primarily due to the reduction of the n-well surface concentration. Figure 7.11b shows the fall of N_{SUB} with n-well drive in time. This correlates well with the results in figure 7.11a.

7.4 Parametric Relationships

Figure 7.12a shows U_0 the MOS surface mobility, plotted against β_F, the forward gain of the vertical bipolar transistor, for each part of the n-well drive in split. From this it can be seen that U_0 decreases as β_F increases. As the well drive-in time is reduced N_D,, the surface concentration of the n-well, will increase. This reduces the surface

mobility. The depth of the n-well also decreases and this makes the effective base width for the vertical bipolar smaller and increases β_F. However it should be noted that the decrease in U_0 is around 16% while the increase in β_F is around 100%.

Figure 7.12b shows the variation of V_{TO} with β_F,. This relationship results for the same reasons discussed above. Figure 7.12c illustrates a very linear relationship between N_{SUB} extracted from the MOS devices vs β_F and this supports the theory of the above relationships.

It can be seen from these relationships that the behaviour of the bipolar parasitic is very closely related to that of the CMOS devices fabricated in the same process. It is this close relationship that will allow bipolar test structures to act as a process control tool for CMOS processes. In the next section bipolar parasitic transistors are used to evaluate CMOS process uniformity. Parts of this work have been presented at an international conference [9] and published in an IEEE journal [10].

7.5 A Process Uniformity Evaluation

By wafer mapping the parameters extracted from the MOS and parasitic devices, the variation of their characteristics due to process non-uniformity can be evaluated. Figure 7.13a and 7.13b show the MOS parameters V_{TO} and μ_o mapped out for wafer 10. These show an obvious variation from the bottom left hand corner of the wafer to the top right. If these maps are compared with Figure 7.13c and 7.13d a similar pattern can be seen in the parasitic bipolar parameter variation. This pattern can also be discerned in Figure 7.13e, a map of parasitic JFET V_T. Figure 7.14 shows a scatter plot of MOS V_{TO} versus the reverse Early voltage extracted from a vertical bipolar transistor. A linear relationship between the two parameters can be observed. Table 7.6 gives the statistical correlation coefficients for these and a number of other parameters. The value of the correlation coefficient represents the extent of the linear association between two variables. These relationships would suggest that the parasitic bipolar parameters are indeed sensitive to the CMOS process variations which have affected the MOS transistors.

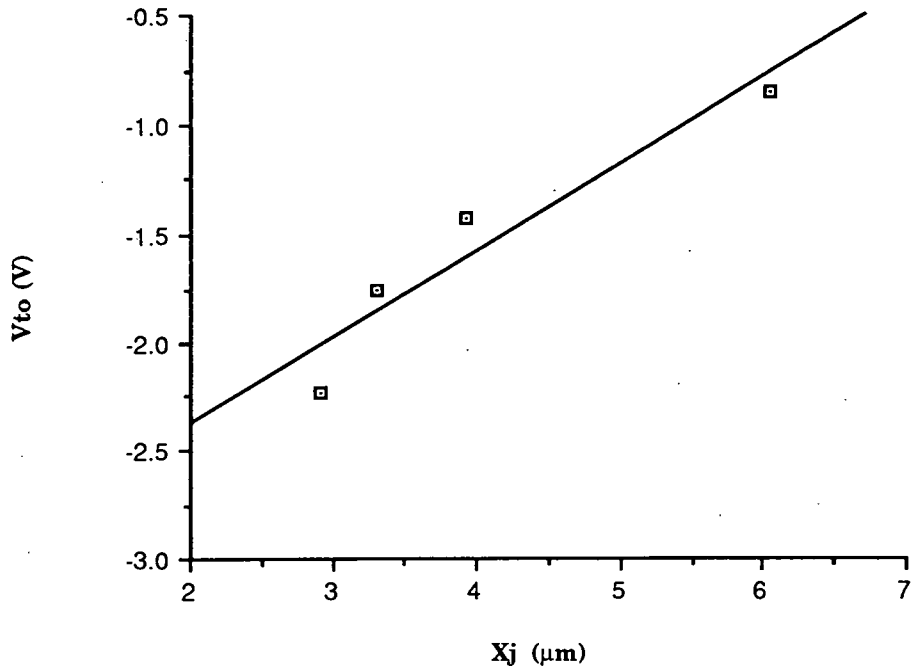


Figure 7.11a p-channel MOS V_{TO} plotted against n-well junction depth.

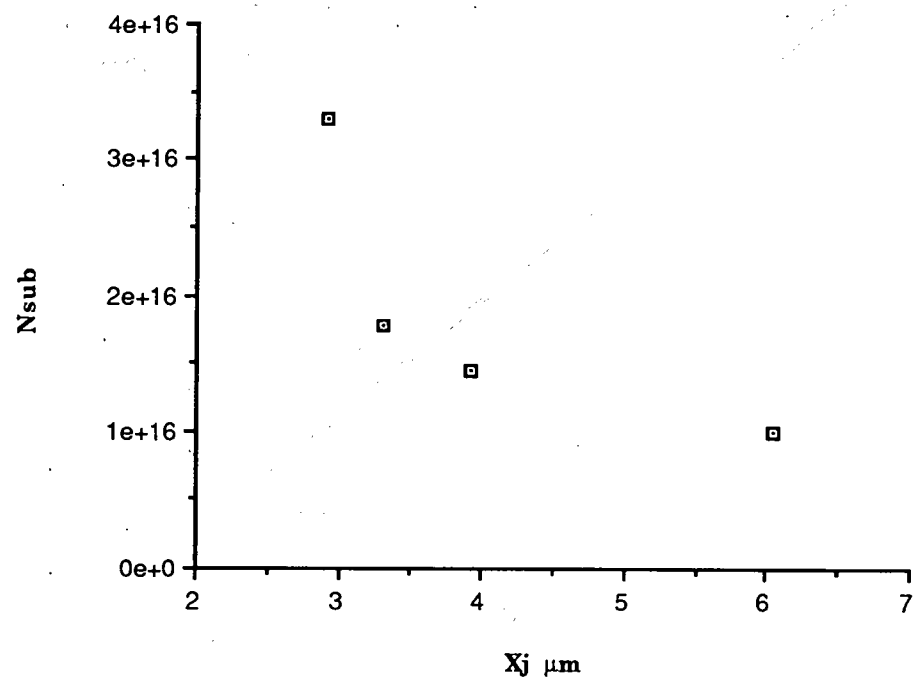


Figure 7.11b p-channel MOS parameter N_{SUB} plotted against n-well junction depth.

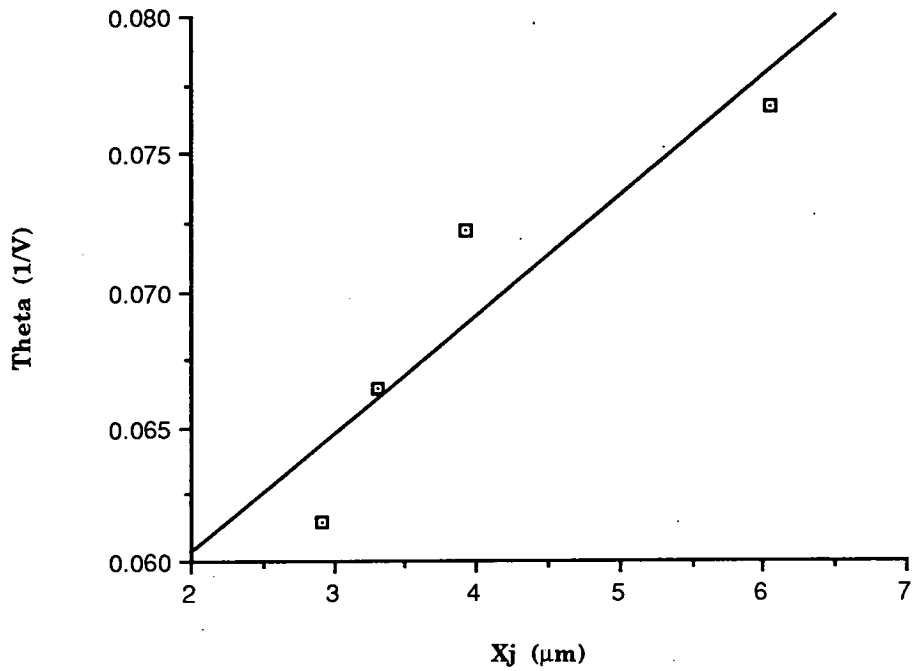


Figure 7.11c p-channel MOS parameter Theta plotted against n-well junction depth.

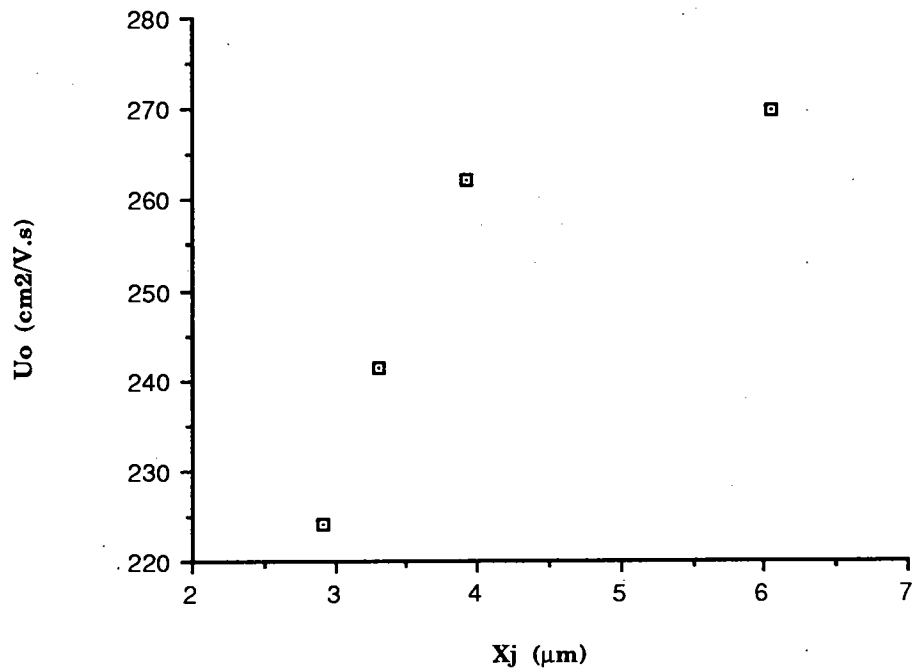


Figure 7. 11d p-channel MOS parameter U_o plotted against n-well junction depth.

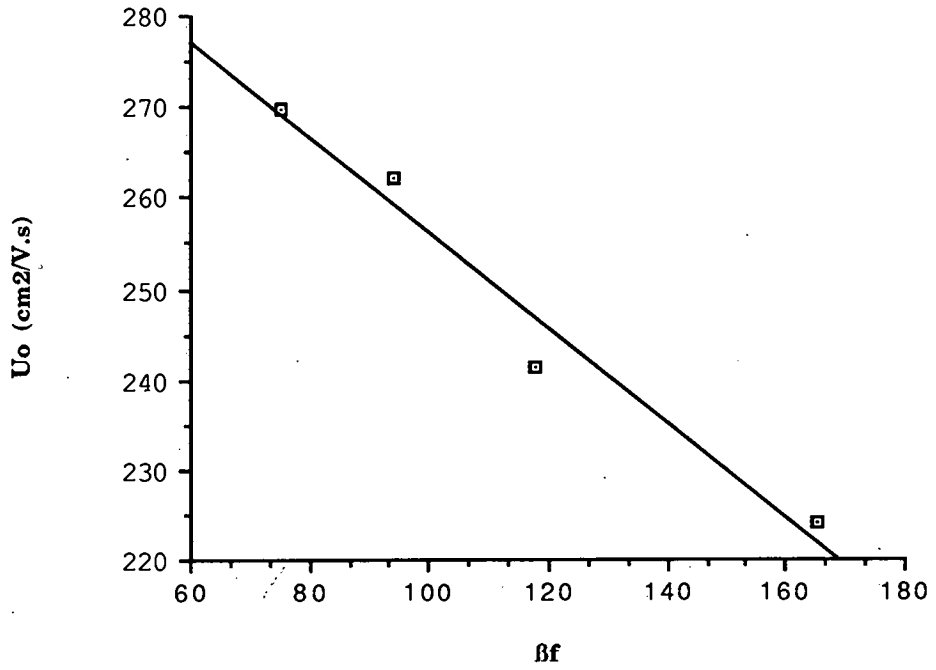


Figure 7.12a p-channel MOS parameter U_O plotted against vertical parasitic bipolar parameter β_F.

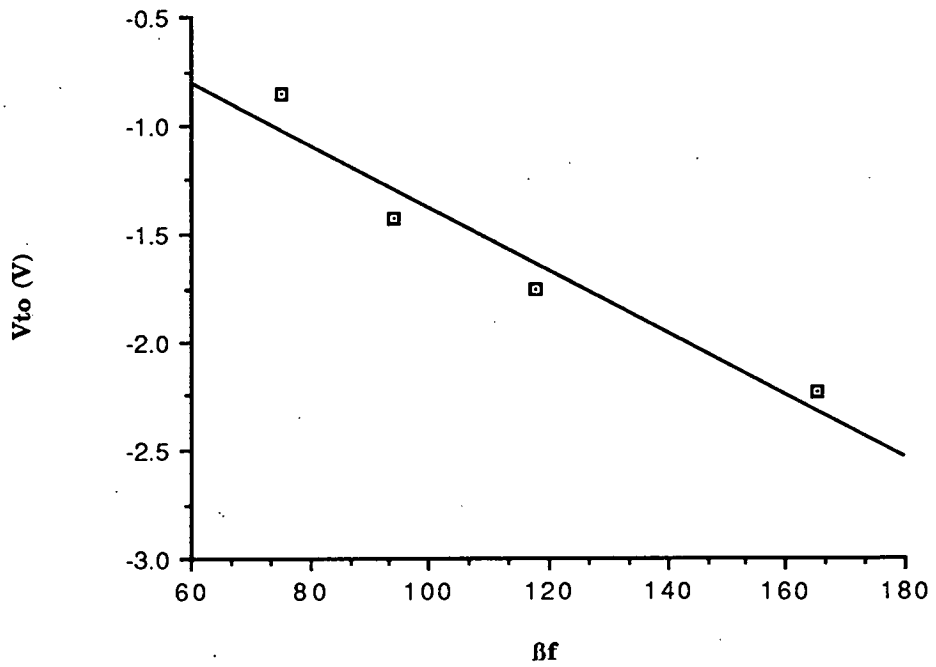


Figure 7.12b p-channel MOS parameter V_{TO} plotted against vertical parasitic bipolar parameter β_F.

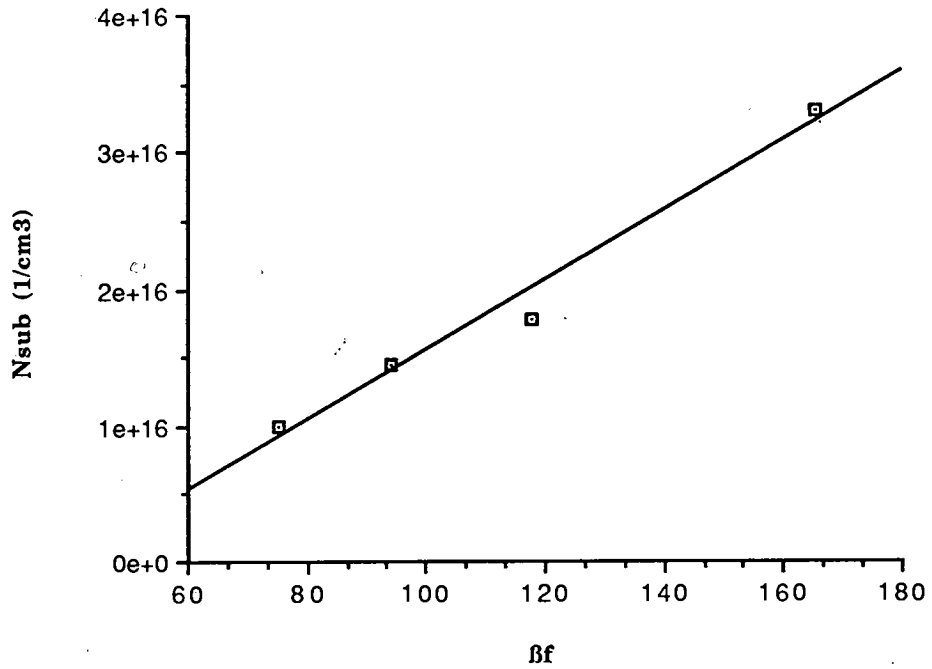


Figure 7.12c p-channel MOS parameter N_{SUB} plotted against vertical parasitic bipolar parameter β_f.

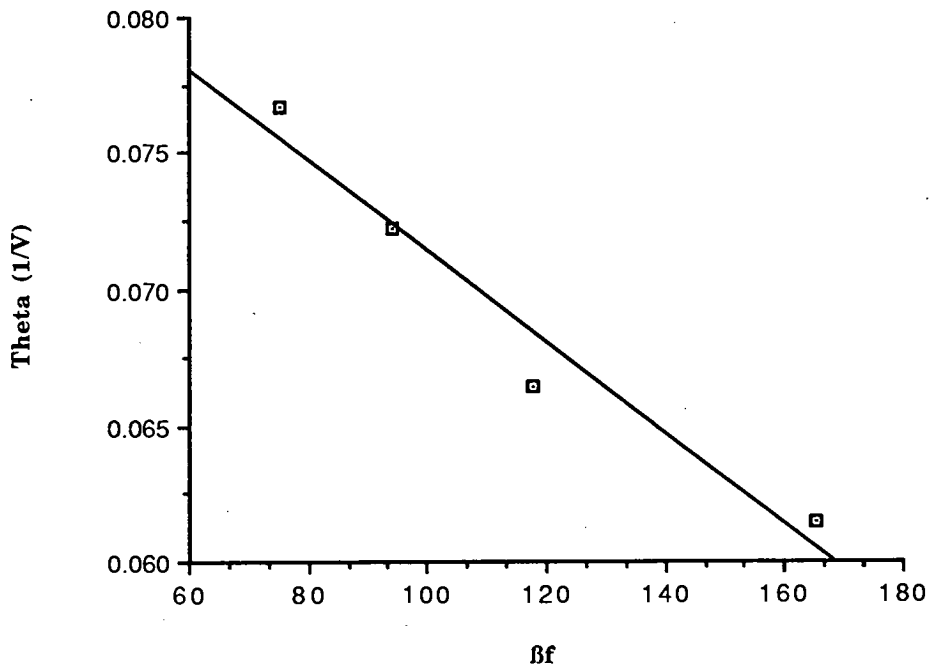


Figure 7.12d p-channel MOS parameter Theta plotted against vertical parasitic bipolar parameter β_f.

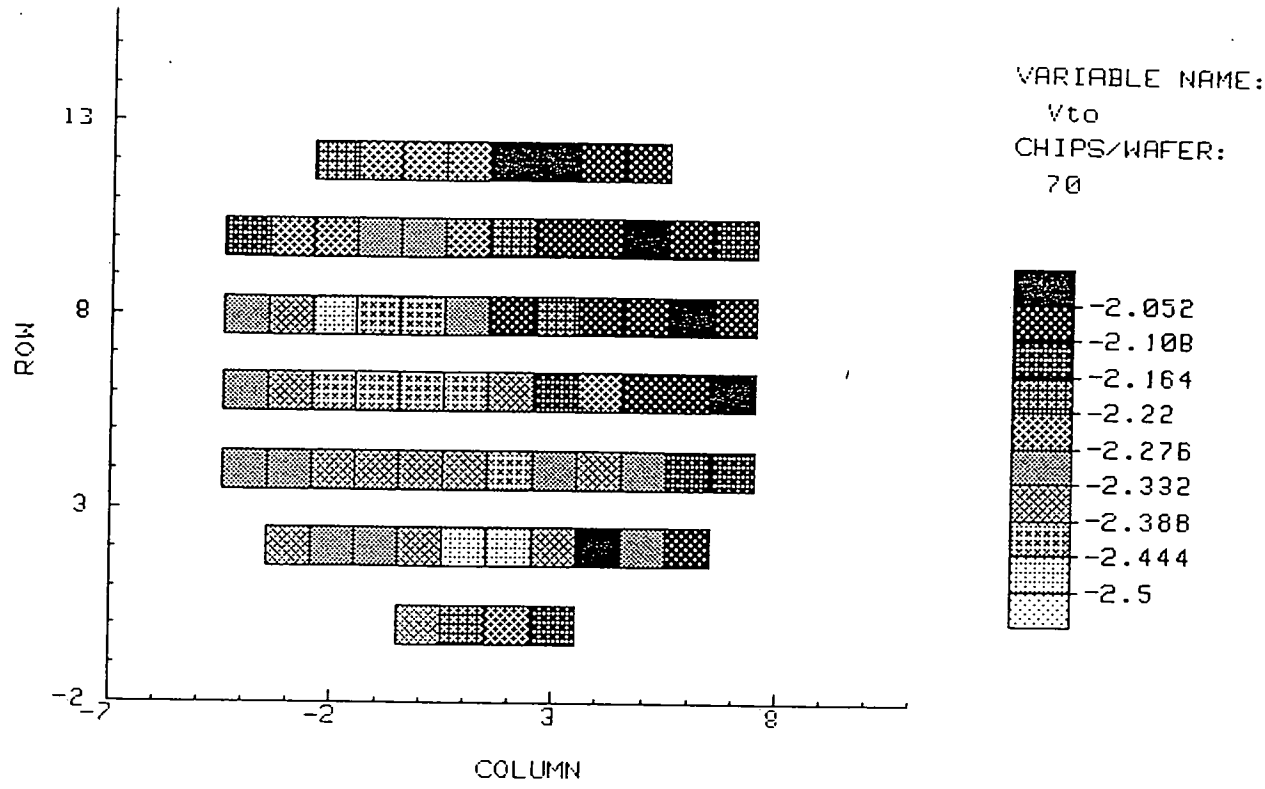


Figure 7.13a Wafer map of V_{TO} for wafer 10.

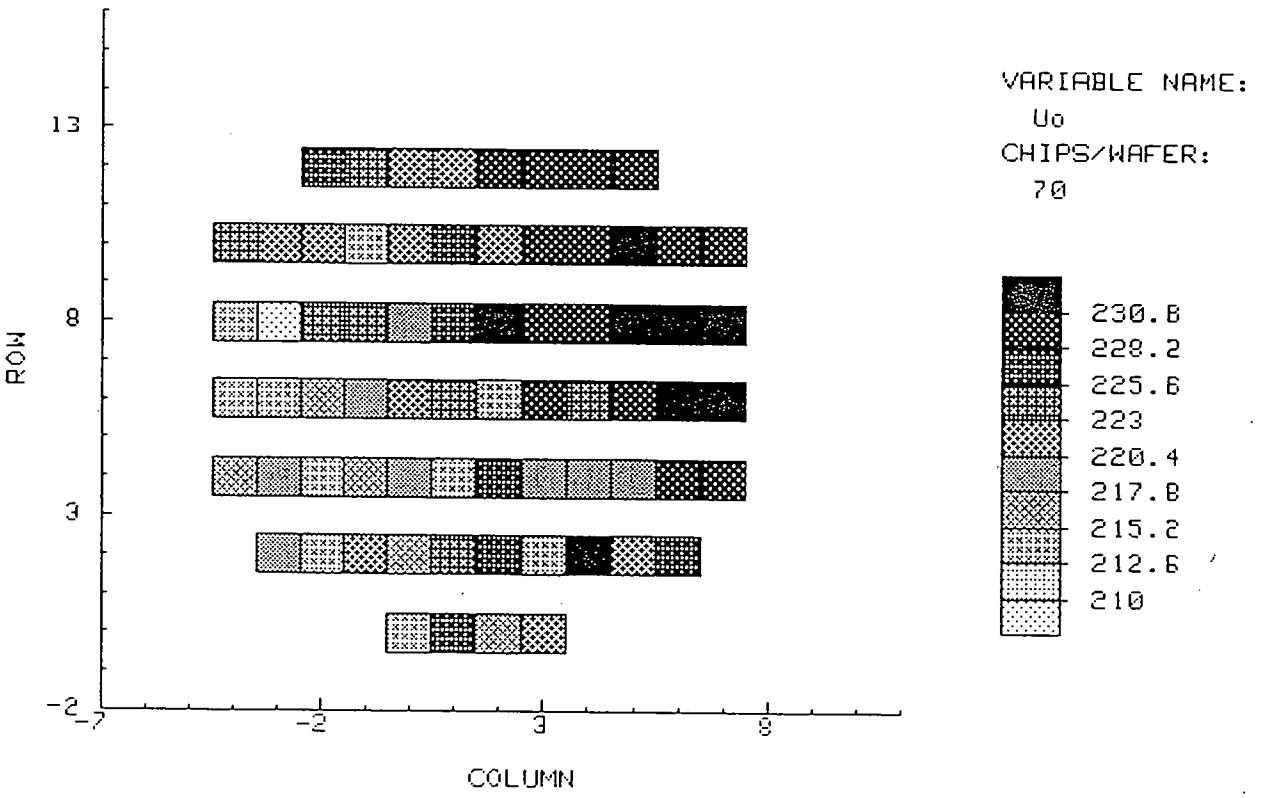


Figure 7.13b Wafer map of μ_o for wafer 10.

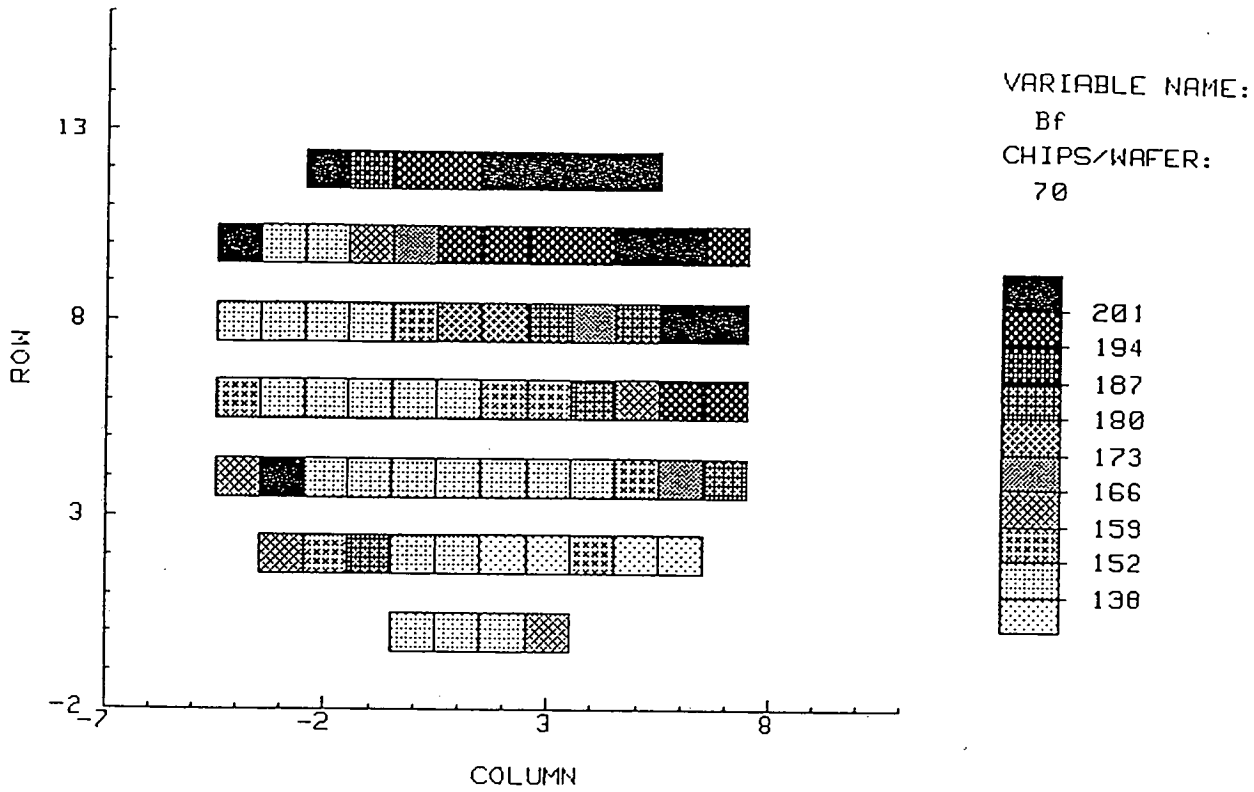


Figure 7.13c Wafer map of β_F for wafer 10.

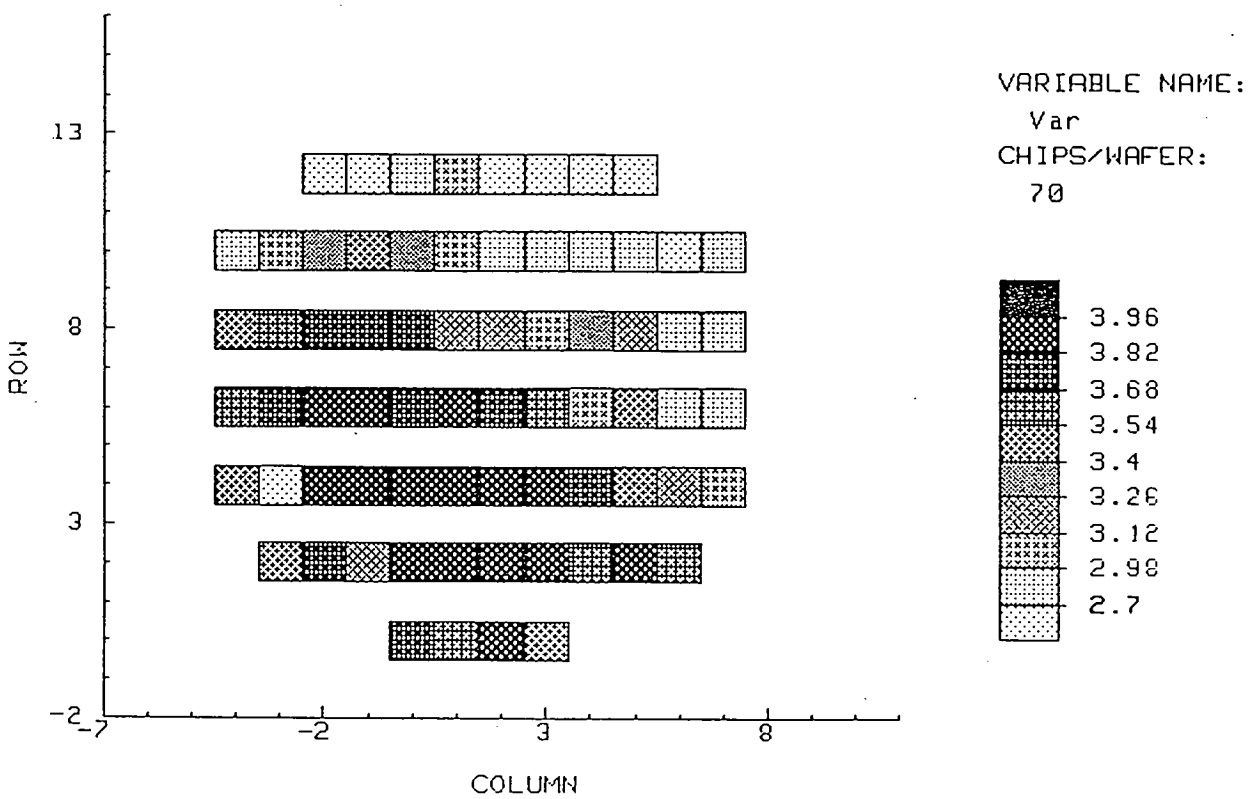


Figure 7.13d Wafer map of V_{AR} for wafer 10.

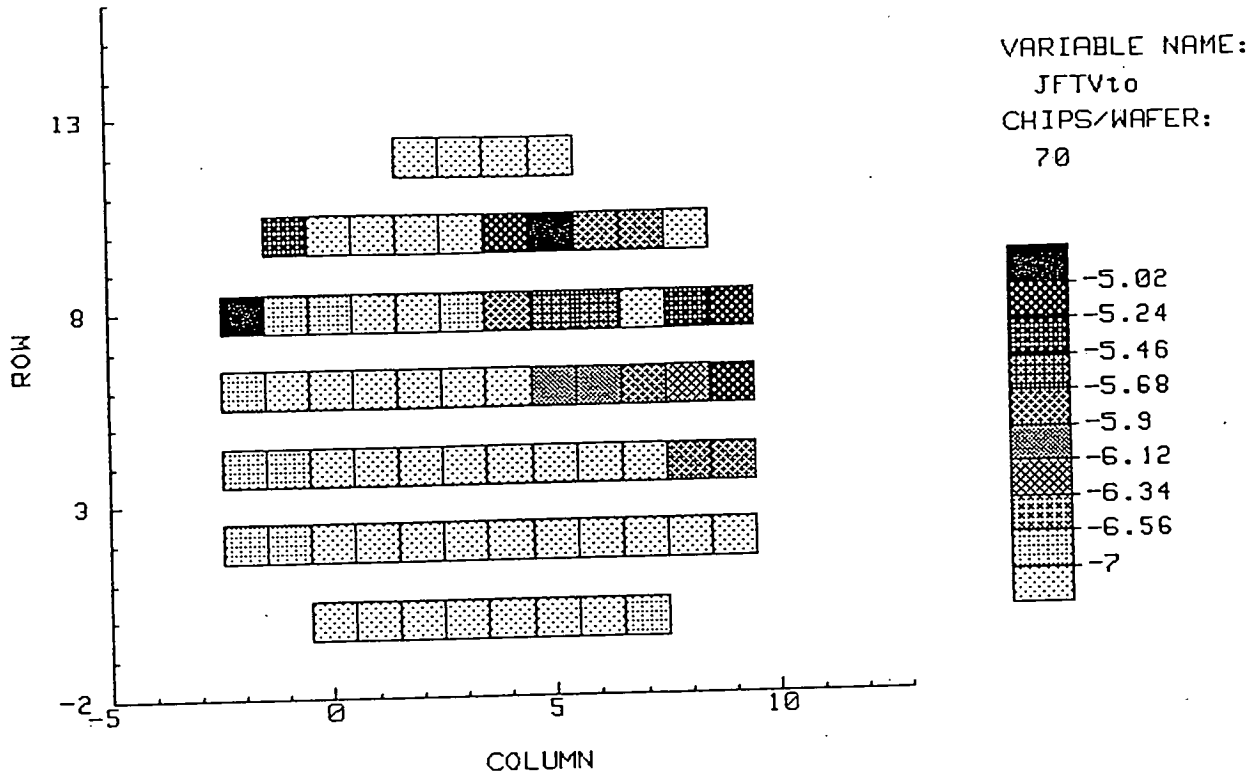


Figure 7.13e Wafer map of JFETV_T for wafer 10.

Parameter Correlation Coefficients from Wafer 10						
	Bipolar				MOS	
	β_F	V_{AF}	V_{AR}	μ_o	V_T	N_{SUB}
I_S	0.900	-0.934	-0.963	0.594	0.690	-0.406
β_F		-0.864	-0.906	0.582	0.689	-0.376
V_{AF}			0.956	-0.566	-0.704	-0.375
V_{AR}				-0.635	-0.749	0.432
μ_o					0.785	-0.210
V_T						-0.614

Table 7.6 Parameter Correlation Coefficients from Wafer 10.

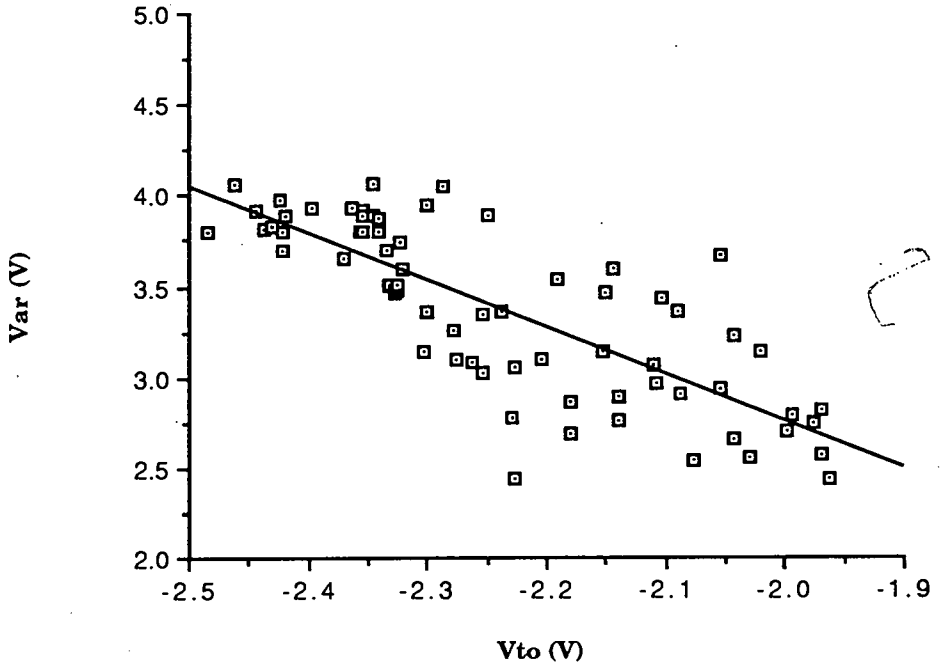


Figure 7.14 Scatter plot of V_{TO} versus V_{AR} for wafer 10.

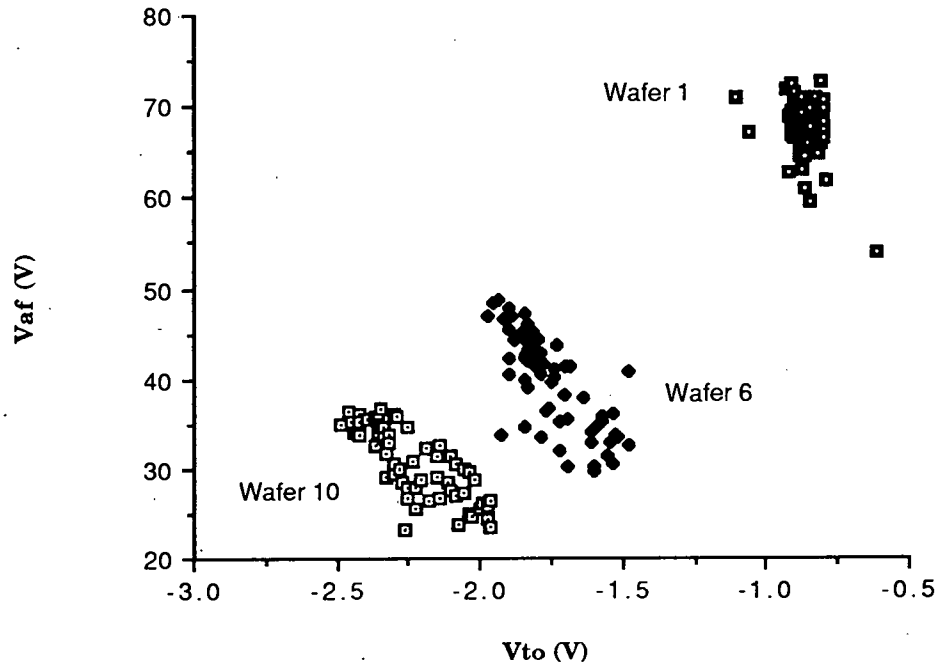


Figure 7.15 Scatter plots of V_{TO} versus V_{AF} for wafers 1,6, and 10.

Figure 7.15 shows the MOS transistor parameter V_{TO} plotted against the forward Early voltage, V_{AF}, for wafers 1,5 and 10. This shows that the early voltage is reduced with decreasing well depth and increasing $|V_{TO}|$. Wafer 10 has the shallowest well depth and exhibits the greatest variation across the wafer as shown in figure 7.16. Some possible causes for this are discussed later.

Equations 7.1-7.4 give theoretical expressions for three of the parameters extracted in the characterisation, I_S, V_{TO}, and β_F.

$$I_S = \frac{qn_i^2 D_{pB} A_E}{Q_B} \tag{7.1}$$

where

$$Q_B = \int_{X_o}^{X_B} N_D(x)dx \tag{7.2}$$

$$\beta_F = \frac{D_{nB} n_{oB}/X_B}{D_{pE} p_{oE}/L_{pE}}. \tag{7.3}$$

Q_B is the Gummel number and N_D is the doping in the base.

$$V_T = V_{FB} + 2\psi_S - \frac{\sqrt{2\varepsilon q N_D(2\psi_B)}}{C_{ox}} \tag{7.4}$$

By considering these equations we can begin to obtain a qualitative feel for the relationship between the MOS and the parasitic devices parameters. This approach can be used to provide extra information which may be required to help isolate the most probable cause of parameter variation. The suspected cause of the variation displayed in Fig. 7.13(a)-(e) is non-uniform well doping which could be due to either implant dose variation or bulk wafer non-uniformity. Uneven dose distribution across the wafer causes a variation in the effective well surface concentration N_D. As N_D decreases then the net charge in the base (Q_B) decreases and, from equation 7.1, I_s will increase. The value N_D for the pMOS transistor is effectively N_{SUB},. Hence, as N_D reduces, $|V_{TO}|$ will fall. The relationship between I_S and $V_{TO,}$ which supports the

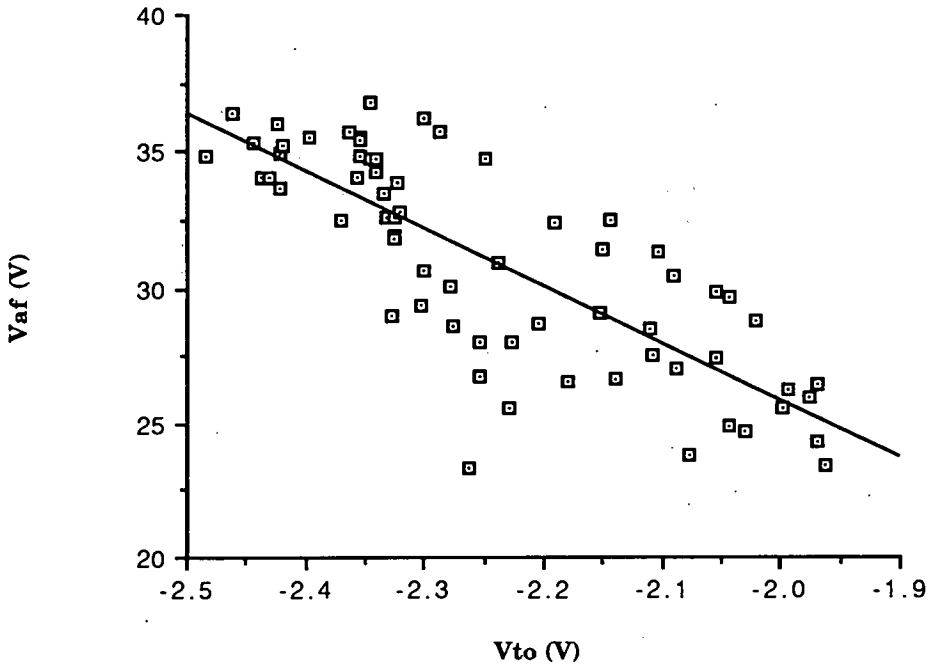


Figure 7.16 Scatter plot of V_{TO} versus V_{AF} for wafer 10.

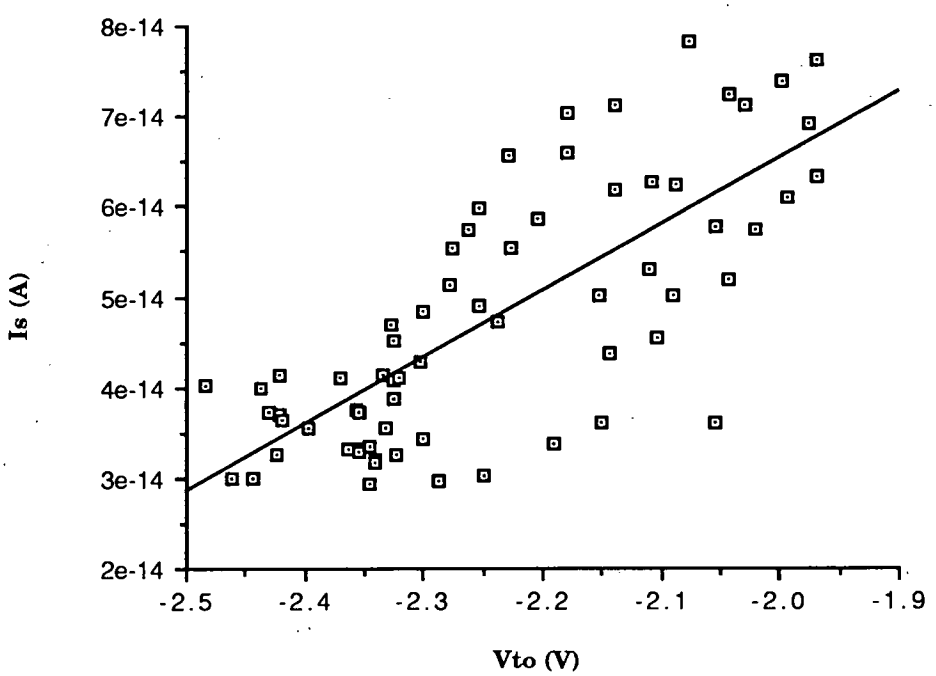


Figure 7.17 Scatter plot of V_{TO} versus I_S for wafer 10.

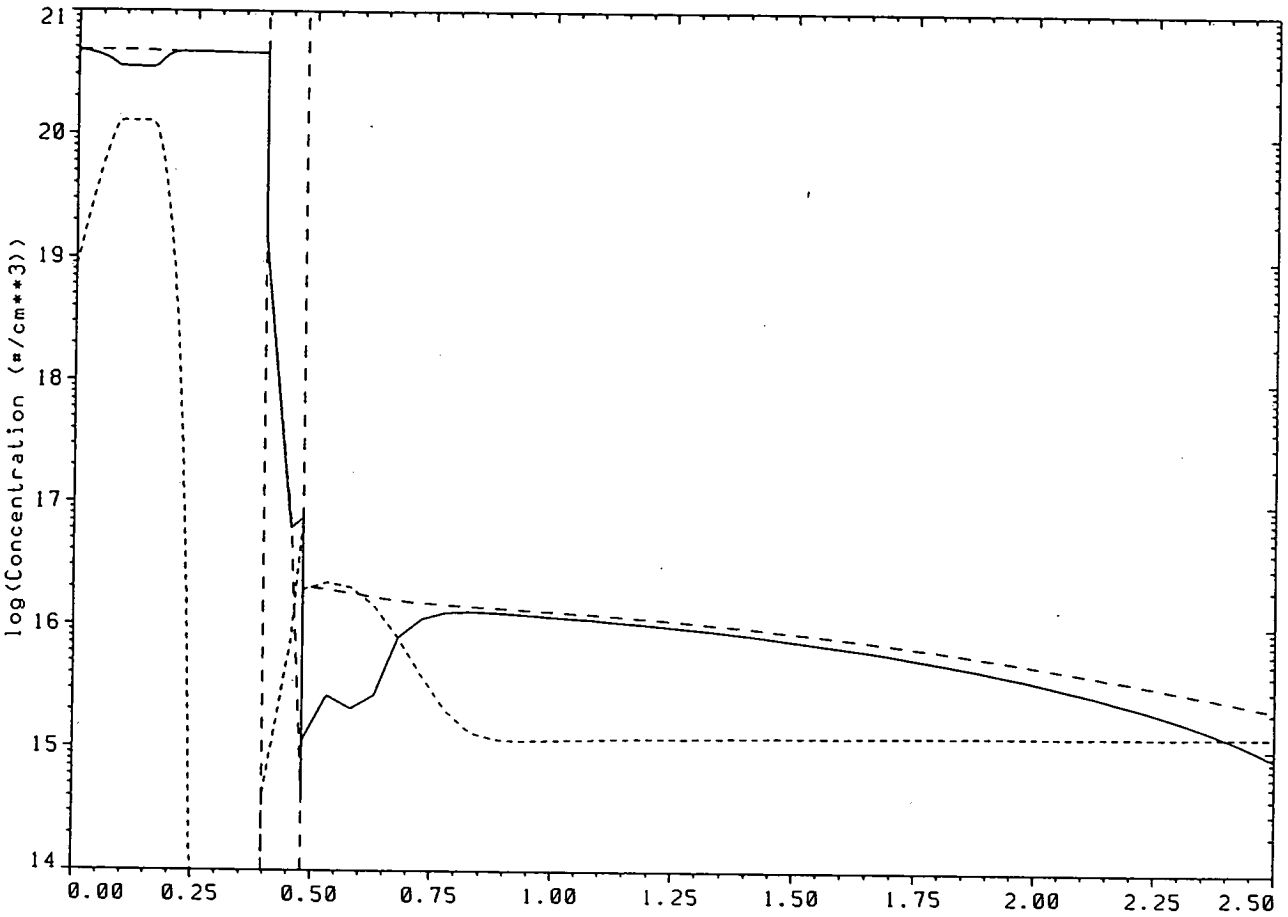


Figure 7.18 SUPREM3 plot of doping under the gate of MOS transistor. This plot results from the simulated processing conditions of wafer 10.

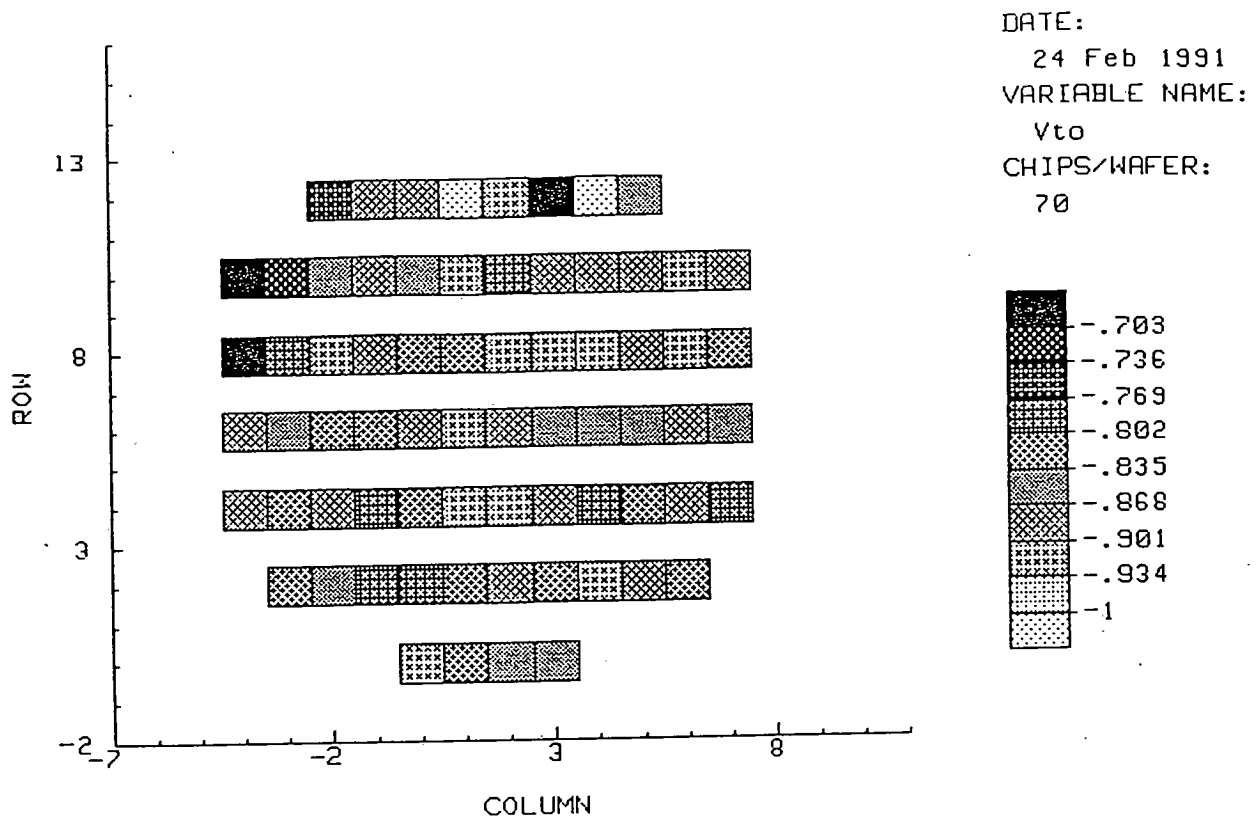


Figure 7.19a Wafer map of V_{TO} for wafer 1.

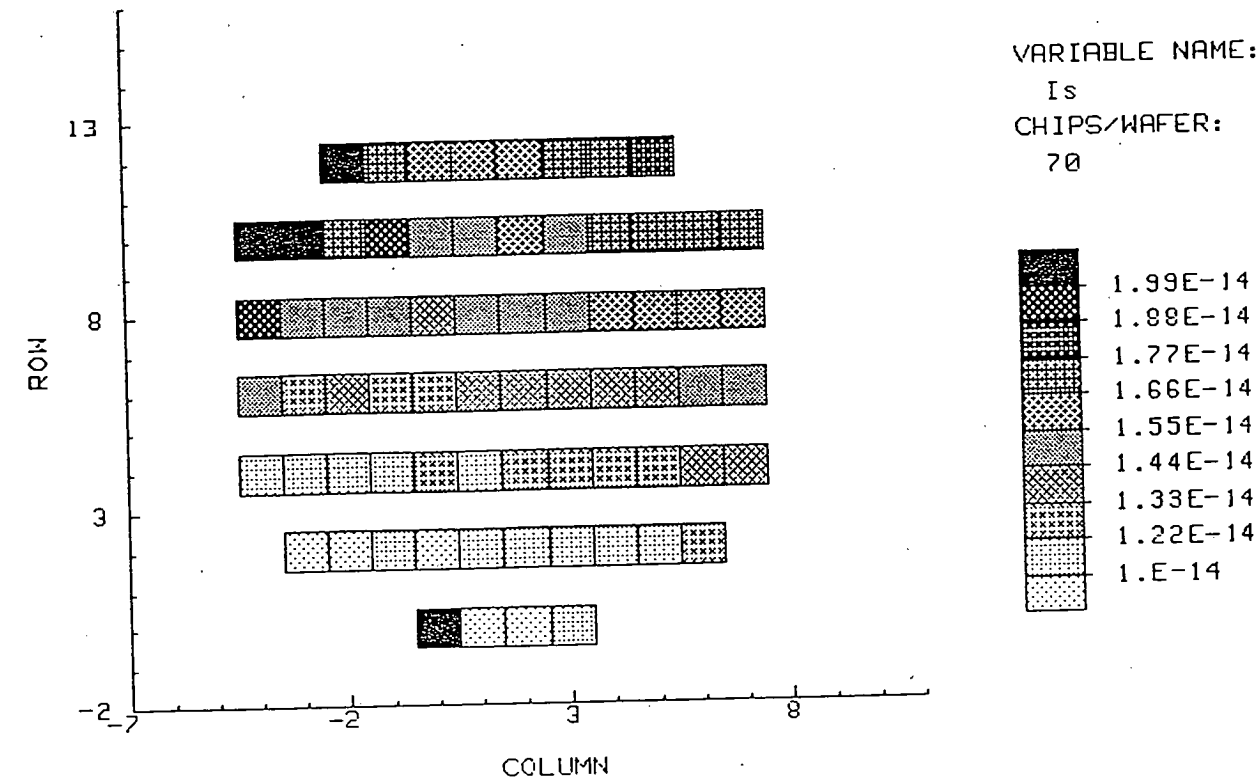


Figure 7.19b Wafer map of I_S for wafer 1.

above conclusion, is illustrated in figure 7.17. It should be noted that as N_D falls the emitter efficiency will be increased while the effective base width (W_B) reduces as the emitter-base junction moves further into the well and the well depth decreases. This accounts for the increase in β_F and the decrease of both V_{AR} and V_T for the JFET. The variations of MOS parameters such as μ_o and N_{SUB} also support this explanation.

The correlations in the above example are enhanced due to N_D being of the same order as the threshold adjust implant. This is illustrated by a SUPREM simulation shown in figure 7.18. In the standard CMOS process the threshold voltage would be less sensitive. Figure 7.19(a) gives the wafer map of V_{TO} for the pMOS transistors of wafer 1 and shows that no pattern can be discerned for the standard well drive-in. However, when the bipolar saturation current I_S is mapped as in figure 7.19(b) we can see a similar pattern to that of wafer 10, shown in figure 7.13(c). This would suggest that the parameters extracted from the bipolar transistors are more sensitive to CMOS process variation than the MOS parameters themselves. This increased sensitivity may give early warning to process shift or variability before the CMOS yield is seriously effected.

In the production environment realistic parameter evaluation for process control must be fast and the test devices small and reliable. The devices presented above fall into that category. The example presented above has used a fairly complex characterisation and extraction package. However the method can be incorporated easily into most test systems by choosing to extract parameters such as β_F at fixed dc bias and V_{AF}. Extraction of these parameters returns a wealth of extra process information for very little expense.

7.6 Conclusion

It has been shown that parasitic transistors can be a very sensitive guide to nonuniformities which are present in a CMOS process. The parasitic transistors presented can be simply included either as scribe channel devices or in a PCC to provide more detailed information about doping concentrations and junction depths. The measurements on these devices can be used to provide additional

information to help determine process problems that are not apparent from the measurement of CMOS devices alone.

References

[1] P.M. Zeitzoff et al,"An Isolated Vertical n-p-n Transistor in an n-Well CMOS Process,"*IEEE journal of solid-state circuits*, vol. SC-20, No. 2, pp. 489-493, 1985.

[2] J.D. Morse and D.H. Navon,"Optimized Design of a Merged Bipolar MOSFET Device,"*IEEE Trans. Electron Devices*, Vol. ED-32, No. 11, pp 2277-2281, 1985.

[3] E.A. Vittoz and O. Neyroud,"A low-voltage CMOS Bandgap Reference,"*IEEE journal of solid-state circuits*, vol. SC-14, No. 3, pp. 573-577, 1979.

[4] E.A. Vittoz,"The design of high-performance Analog Circuits on Digital CMOS Chips,"*IEEE journal of solid-state circuits*, vol. SC-20, No. 3, pp. 657-665, 1985.

[5] R. Ching-Yuh Fang and J.L. Moll,"Latchup model for the Parasitic p-n-p-n Path in Bulk CMOS,"*IEEE Trans. Electron Devices*, Vol. ED-31, No. 1, pp 113-120, 1984.

[6] P.V. Dressendorfer and A. Ochoa,"An Analysis of the modes of operation of Parasitic SCR's,"*IEEE Trans. Nuclear Science*, Vol. NS-28, No. 6, pp.4288-4294, 1981.

[7] G.J. Hu,"A Better Understanding of CMOS Latch-Up,"*IEEE Trans. Electron Devices*, Vol. ED-31, No. 1, pp 62-67, 1984.

[8] D.S. Perloff, F.E. Wahl and J.D. Reimer,"Contour maps reveal non-uniformity in semiconductor processing,"*Solid State Technology*, pp.30-42, 1977.

[9] D.Wilson, J. M. Robertson, R. Holwill and A. J. Walton,"CMOS

Process Uniformity Evaluation," *Proc. ICMTS*, Long Beach Calif, 1989.

[10] D.Wilson, A. J. Walton, J. M. Robertson and R. Holwill,"Characterization of Parasitic Transistors to Evaluate CMOS Process Uniformity," *IEEE Trans. Semiconductor Manufacturing*, Vol. 4. No. 3, pp 241-249, 1991.

Chapter 8

Conclusions and Further Work

8.1 Conclusions

CMOS has become the dominant technology in integrated circuit manufacture. Current market observations show that it will continue to dominate and expand (chapter 1). BiCMOS will also expand and is predicted to attain a market share of at least 5% by 1994. In electronics like all manufacturing technologies the need for quality control has grown significantly. The electronics industry in particular with many devices in critical applications needs to assure it's customers of their products quality and reliability. Quality and reliability now ranks alongside price and new product innovation in most electronics company's marketing strategy. Statistical process control (SPC) has grown to meet the need for quality assurance (QA). Most companies now have their own QA policies which have at their heart a commitment to SPC.

SPC needs uncompromising metrics to be able to successfully identify and eliminate non-conforming material. In the particular case of CMOS, critical dimensions of 1.0 - 0.8μm are now commonplace and 0.7-0.6μm devices are emerging from development to production. New metrics are now required to control device aspects which previously had no discernable or adverse effect to device performance.

This thesis has looked at this problem and has defined some new techniques and metrics for the process control of a standard CMOS process. The approach has been a novel one. The parasitic transistors available in a CMOS process are characterised and test structures examined to assess their ability to provide parametrics which can be

used for CMOS process control. A review has been presented of CMOS technology's past and current status. Introductory theory of operation of bipolar and MOS transistors has been presented to enable the understanding of the relationship between parasitic bipolar transistors and MOS transistors fabricated in a standard CMOS process. The role of SPC has been discussed and a case study of Motorola's six sigma policy has been presented (chapter 3). The design and fabrication of test structures suitable for the objective of the characterisation was presented in chapter 4. Chapters 5,6 and 7 offer techniques for process control which can be applied to any standard CMOS process. The techniques presented include the following:

1. The characterisation of parasitic JFETs to provide well depth information electrically. This technique avoids the use of expensive and time consuming methods, such as Secondary Ion Mass Spectroscopy (SIMS), and the traditional bevel and stain approach to junction profiling.

2. The use of parasitic lateral bipolar transistors to estimate the sideways diffusion component (ΔL) associated with MOS transistors fabricated in a CMOS process. The technique presented is a quick three point measurement procedure which gives comparable results to conventional methods. The technique avoids the inaccuracies of short channel effects on the transconductances of sub-micron MOS devices.

3. The use of parasitic bipolar test structures to evaluate CMOS process uniformity. This technique was evaluated through the examination of the relationship between bipolar parameters and MOS parameters extracted from devices fabricated on the same chip. Wafer mapping enabled the wafer-scale comparison of each device's ability to reflect non-uniform CMOS processing through the examination of their electrical characteristics. It was found that some of the bipolar transistors were more sensitive to CMOS processing than

those of the CMOS devices.

All of the devices discussed, designed and fabricated for this thesis were fabricated using a standard CMOS process. No extra mask steps or processing steps were needed. A consequence of this is that they can be incorporated into any CMOS test chip or scribe grid structure. There are two main benefits to be gained by using test structures like those described in this thesis:

1. The test structures give a wealth of extra process monitoring parameters which are not routinely extracted from a MOS process.

2. The analysis of these devices will provide an introductory knowledge of bipolar device operation to process/device engineers before this becomes an absolute necessity in the production of BiCMOS circuits.

8.2 Further Work

A number of ideas are presented in this thesis which require some further work. The experiments presented were designed and fabricated with a 5μm CMOS process (albeit with step and repeat capability). Although the basic structures would be the same they could be re-designed with smaller geometry design rules. The test structures like most traditional MOS structures would be scaled suitable for the process. However, some of the ideas presented require some more characterisation work:

1. The concept of using lateral bipolar transistors to estimate ΔL could be applied to lateral devices formed by a CMOS process with a Lightly Doped Drain (LDD) transistor structure. The simple model defined in chapter 6 may not apply directly to structures formed in this manner. However it is likely that lateral transport would occur primarily at the edge of the LDD (shortest base width) and the model could be modified to

fit the physical transport of the devices.As short channel effects become more pronounced this method provides a relatively simple method for extracting ΔL.

2. It was stated in chapter 4 that although an n-well process was used, the approach would be just as valid in p-well or twin well technology. Each pnp device presented in the thesis has an equivalent npn structure which could be characterised.

3. A complete test vehicle could be designed which would contain all relevant test structures for the process control of a CMOS or BiCMOS process.

4. The ability of parasitic JFETs to monitor well junction depth could be assessed using SIMS profiling rather than relying on simulations to estimate the method's accuracy.

5. The process uniformity evaluation presented in chapter 7 compared DC bipolar parameters with traditional MOS parameters. With CMOS critical dimensions being pushed towards the sub-half micron level, AC characteristics of bipolar transistors fabricated in a CMOS process may provide useful process control information. Unfortunately, the hardware required for AC characterisation of bipolar transistors remains expensive in comparison with that required for DC analysis.

Parasitic transistors fabricated in a CMOS process provide useful parameters for process control. These parameters have even been shown, in some cases, to be more sensitive to CMOS process non-uniformities than those extracted from MOS devices. The direction of integrated circuit manufacturing technologies is moving inevitably towards the mixed approach of BiCMOS. One may look forward to BiCMOS becoming the dominant technology some time in the next century. The adoption of the concepts presented in this thesis will

provide valid process control information for today's CMOS processes and an insight into the control of future BiCMOS processes.